A Stochastic Model For Critical Illness Insurance

Erengul Ozkok

Submitted for the degree of

Doctor of Philosophy

on completion of research in the

Department of Actuarial Mathematics & Statistics,

School of Mathematical and Computer Sciences,

Heriot-Watt University

May 2011

I hereby declare that the work presented in this thesis was carried out by myself at Heriot-Watt University, Edinburgh, except where due acknowledgement is made, and has not been submitted for any other degree.

Erengul Ozkok (Candidate)

Dr George Streftaris (Supervisor)

Professor Howard R. Waters (Supervisor)

Professor A. David Wilkie (Supervisor)

Date

Abstract

In this thesis, we present methods and results for the estimation of diagnosis inception rates for Critical Illness Insurance (CII) claims in the UK by cause. This is the first study which provides a stochastic model for the diagnosis inception rates for CII. The data are supplied by the UK Continuous Mortality Investigation and relate to claims settled in the years 1999 - 2005. First, we develop a model for the delay between dates of diagnosis and settlement of claims in CII using a generalised-linear-type model with Burr errors under both Bayesian and maximum likelihood approach. Variable selection using Bayesian methodology to obtain the best model with different prior distribution setups for the parameters is applied. For comparison purposes, a lognormal model and frequency-based model selection techniques are also considered. The non-recorded dates of diagnosis and settlement have been included in the analysis as missing values using their posterior predictive distribution and Markov Chain Monte Carlo methodology. Missing dates of diagnosis are estimated using the parsimonious claim delay distribution. With this complete data set, diagnosis inception rates for all causes (combined) and for specific causes are estimated using an appropriate claim delay distribution where the observed numbers of claim counts are assumed to have a Poisson distribution. To model the crude rates, a generalised linear model with Poisson errors and log-link function is used.

Acknowledgements

I am extremely grateful to my supervisors Dr George Streftaris, Professor Howard Waters and Professor David Wilkie for their vast amount of time, guidance, support, encouragement and patience. I could not have had better supervision; their doors were always open to me. I am indebted to them and very honoured to be their student.

I would like to thank the Continuous Mortality Investigation of the Institute and Faculty of Actuaries for providing the data which is used in the thesis and for quick responses whenever we need them.

I thank Hacettepe University for being a sponsor for my PhD.

I express my appreciation to Dr Ioannis Ntzoufras and Dr Dimitris Fouskakis for helping me to improve this research by their comments.

I am also very grateful to all my friends who made my life brighter here in Edinburgh.

I wish to express my deepest gratitude to my family. Without their encouragement and support I could not have had this excellent experience.

Contents

Abstract **i**

Acknowledgement **ii**

1 Introduction **1**

1.1 Critical Illness Insurance 1

1.2 Overview of past results 3

1.3 Stochastic modelling for inception rates 6

1.4 Outline of the thesis 10

2 Data **12**

3 Modelling CDD I: Without considering the missing values and growth rates **20**

3.1 Introduction 20

3.2 The observed delay 21

3.3 Modelling under the assumption of no growth within offices 23

3.3.1 The null model 26

3.3.2 Analysis with covariates 30

3.4 Prediction under the two models 41

4 Modelling CDD II: Without considering the missing values and taking growth rates into account **44**

4.1 Introduction 44

4.2 Modelling when the business growth is taken into account 47

4.3 Prediction 57

5 Selection of claim-specific covariates **61**

5.1 Introduction 61

5.2 Selection of claim-specific covariates without growth factor 62

5.2.1 Gibbs variable selection (GVS) 62

5.2.2 Variable selection using marginal likelihoods 69

5.2.3 Maximum likelihood based methods 74

5.3 Selection of claim-specific covariates with growth factor 80

6 Modelling CDD III: Including the missing values **86**

6.1 Introduction 86

6.2 Including the missing values and assuming no growth within offices . 87

6.3 Considering the missing values and the growth rate 92

7 Diagnosis inception rates I: All-cause rates **100**

7.1 Introduction 100

7.2 Modelling techniques 101

7.3 All-cause rates without restriction 106

7.4 All-cause rates including the CMI variables 117

7.5 Sensitivity of the inception rates to delay estimates 135

7.5.1 Sensitivity analysis 135

8 Diagnosis inception rates II: Cause-specific rates **145**

8.1 Structure of the model 146

8.2 Best models for specific causes 147

8.3 Comparison of cause-specific rates with the all-causes rates 192

9 Conclusions and Further Research **203**

9.1 Conclusions 203

9.2 Further research 207

References **209**

List of Tables

2.1 Details of the submission groups. 14
2.2 Number of offices and number of sales channels. 14
2.3 Average observed delays between dates of diagnosis, notification, admission and settlement (in days). 16
2.4 Number of claims and percentages by various factors. 16
2.5 Grouping cause of claim. 19

3.1 Posterior and ML estimation under the null model. 28
3.2 Information criteria under the null model. 28
3.3 Definitions of the covariates. 30
3.4 Coefficients of the Burr model without growth rate. 33
3.5 Coefficients of the LN model without growth rate. 37
3.6 Values of information criteria of the models without growth rate. . . 39
3.7 Scenarios for prediction of the CDD. 42
3.8 Posterior estimates of mean delays for the scenarios in Table 3.7 for the Burr and LN model (days). 43

4.1 Growth rates from the inforce data for offices between successive years. 46
4.2 Coefficients of the Burr model with growth rate. 49
4.3 Coefficients of the LN model with growth rate. 53
4.4 Values of information criteria of the models with growth rate. 56
4.5 Posterior estimates of mean delay for the scenarios in Table 3.7 for the Burr and LN model with growth rate (days). 58

5.1 Parameter inclusion probabilities and model probabilities under the Burr model with independent normal priors. 65

5.2 Parameter inclusion probabilities and model probabilities under the LN model with independent normal priors. 66
5.3 Parameter inclusion probabilities and model probabilities under the Burr model with empirical and Zellner's g-priors. 70
5.4 Parameter inclusion probabilities and model probabilities under the LN model with empirical and Zellner's g-priors. 71
5.5 Exact marginal likelihoods (EML) for the lognormal model. 72
5.6 Laplace Approximation for the Burr Model. 74
5.7 Variable selection with backward stepwise method for the Burr model. 76
5.8 Variable selection with backward stepwise method for the lognormal model. 77
5.9 LRT values for each step given in Table 5.7 and 5.8. 77
5.10 DIC values of the selected models 78
5.11 Estimates of the parameters under the selected Burr model (m_{1013}). 79
5.12 Parameter inclusion probabilities and model probabilities under the Burr model with empirical and Zellner's g-priors when business growth is taken into account. 82
5.13 Comparison of the Burr and LN models with growth rate. 83
5.14 Estimates of the parameters under the best Burr model (m_{981}) without missing values with growth rate. 84

6.1 Posterior estimates of parameters under the selected Burr model with missing values (m_{1013}). 90
6.2 Coefficients of the Burr model (m_{981}) with missing values and with growth rate. 94
6.3 Scenarios for prediction of the CDD under the selected Burr model (m_{981}). 97
6.4 Posterior estimates of the mean of the delay distribution under different scenarios given in Table 6.3 using the selected Burr model (m_{981}) with growth rates. 98

7.1 Definitions of the covariates used in the modelling of the intensity rates. 101
7.2 Selected covariates, log-likelihood values and BIC from fitting different $g_r(x), f_s(x)$ polynomials. 107

7.3 ML estimates of parameters under the best model for all-cause rates. 108
7.4 Selected covariates, log-likelihood values and BIC from fitting different $g_r(x), f_s(x)$ polynomials. 118
7.5 ML estimates of parameters under the best model which includes the CMI variables for all-cause rates. 119
7.6 Selected covariates, log-likelihood values and BIC from fitting different $g_r(x), f_s(x)$ polynomials. 119
7.7 ML estimates of parameters under the model with the CMI variables for all-cause rates. 120

8.1 ML estimates of parameters under the best model for CABG. 149
8.2 ML estimates of parameters under the best model for cancer. 153
8.3 ML estimates of parameters under the best model for death. 159
8.4 ML estimates of parameters under the best model for heart attack. . 166
8.5 ML estimates of parameters under the best model for kidney failure. 169
8.6 ML estimates of parameters under the best model for major organ transplantation. 172
8.7 ML estimates of parameters under the best model for multiple sclerosis. 175
8.8 ML estimates of parameters under the best model for other diseases. 179
8.9 ML estimates of parameters under the best model for stroke. 183
8.10 ML estimates of parameters under the best model for total and permanent disability. 188

List of Figures

1.1 A multi-state model for specific causes for CII. 9

3.1 Histogram of claim settlement delay (in days). 22
3.2 Box plots of observed delay by cause (in days). 22
3.3 CDF of the diagnosis – settlement interval. 29
3.4 Deviance residuals of the Burr model without growth rate. 35
3.5 Deviance residuals of the LN model without growth rate. 38
3.6 Comparison of posterior estimates of the coefficients under Burr (black line) and LN (red line) model. Bars show 95% credible intervals and bullets show posterior means. 40
3.7 Comparison of posterior estimates of the mean delay under different scenarios using the Burr (black line) and LN model (red line). Bars show 95% credible intervals and bullets show posterior means. Vertical lines show the posterior means of the first scenarios under the two models. 43

4.1 Deviance residuals of the Burr model with growth rate. 50
4.2 Comparison of the posterior estimates of the coefficients under the Burr model with and without growth rate. 51
4.3 Comparison of posterior estimates of the coefficients under the LN model with and without growth rate. 54
4.4 Comparison of posterior estimates of the coefficients under the Burr and LN models. 55
4.5 Deviance residuals of the LN model without missing values and with growth rate. 56

4.6 Posterior estimates of the means (days) of the scenarios under the Burr (black line) and LN model (red line) with growth rate. Bars show 95% credible intervals and bullets show posterior means. Vertical lines show the posterior means of the first scenarios under the two models. . . . 59

4.7 Posterior estimates of the means (days) of the scenarios under the Burr and LN model with (red line) and without (black line) growth rate. Bars show 95% credible intervals and bullets show posterior means. Vertical lines show the posterior means of the first scenarios under the two models. 60

6.1 Chronological order of the dates relating to a claim. 88

6.2 Comparison of posterior densities of model parameters ($\beta_1 - \beta_9$) under the selected Burr model (m_{1013}) with (red dashed line) and without (black solid line) missing values. 91

6.3 Comparison of posterior densities of model parameters (β_{10}, α, τ) under the selected Burr model (m_{1013}) with (red dashed line) and without (black solid line) missing values. 92

6.4 Comparison of posterior densities of model parameters ($\beta_1 - \beta_9$) under the selected Burr model (m_{981}) including the growth rate, with (red dashed line) and without (black solid line) missing values. 95

6.5 Comparison of posterior densities of model parameters (β_{10}, α, τ) under the selected Burr model (m_{981}) including the growth rate, with (red dashed line) and without (black solid line) missing values. 96

6.6 Comparison of posterior estimates of the mean delay under different scenarios using the selected Burr model (m_{981}) including (red solid line) and excluding (black solid line) the missing information. Bars show 95% credible intervals and bullets show posterior means. 99

7.1 Observed and expected number of claims. 104

7.2 Male smoker crude inception rates divided by male non-smoker crude inception rates. 107

7.3 Graphs of diagnosis inception rates for non–smokers and durations 0 & 1. 110

7.4 Graphs of diagnosis inception rates for non–smokers and durations 2 & 3. 111
7.5 Graphs of diagnosis inception rates for non–smokers and durations 4 & 5+. 112
7.6 Graphs of diagnosis inception rates for smokers and durations 0 & 1. 113
7.7 Graphs of diagnosis inception rates for smokers and durations 2 & 3. 114
7.8 Graphs of diagnosis inception rates for smokers and durations 4 & 5+. 115
7.9 Comparison of diagnosis inception rates for non–smokers vs smokers under the best model for policy duration 0. 116
7.10 Graphs of diagnosis inception rates for males, non–smokers, full accelerated policies and durations 0 & 1. 122
7.11 Graphs of diagnosis inception rates for males, non–smokers, full accelerated policies and durations 2 & 3. 123
7.12 Graphs of diagnosis inception rates for males, non–smokers, full accelerated policies and durations 4 & 5+. 124
7.13 Graphs of diagnosis inception rates for males, smokers, full accelerated policies and durations 0 & 1. 125
7.14 Graphs of diagnosis inception rates for males, smokers, full accelerated policies and durations 2 & 3. 126
7.15 Graphs of diagnosis inception rates for males, smokers, full accelerated policies and durations 4 & 5+. 127
7.16 Graphs of diagnosis inception rates for females, non–smokers, full accelerated policies and durations 0 & 1. 128
7.17 Graphs of diagnosis inception rates for females, non–smokers, full accelerated policies and durations 2 & 3. 129
7.18 Graphs of diagnosis inception rates for females, non–smokers, full accelerated policies and durations 4 & 5+. 130
7.19 Graphs of diagnosis inception rates for females, smokers, full accelerated policies and durations 0 & 1. 131
7.20 Graphs of diagnosis inception rates for females, smokers, full accelerated policies and durations 2 & 3. 132

7.21 Graphs of diagnosis inception rates for females, smokers, full accelerated policies and durations 4 & 5+. 133
7.22 Comparison of diagnosis inception rates for non–smokers vs smokers using CMI variables for males and duration 0. 134
7.23 Graphs of relative diagnosis inception rates with different missing delay estimates and confidence intervals of the inception rates based on the median of the CDD (as a ratio of inception rates based on the median of the CDD) for non–smokers and durations 0 & 1. 139
7.24 Graphs of relative diagnosis inception rates with different missing delay estimates and confidence intervals of the inception rates based on the median of the CDD (as a ratio of inception rates based on the median of the CDD) for non–smokers and durations 2 & 3. 140
7.25 Graphs of relative diagnosis inception rates with different missing delay estimates and confidence intervals of the inception rates based on the median of the CDD (as a ratio of inception rates based on the median of the CDD) for non–smokers and durations 4 & 5+. 141
7.26 Graphs of relative diagnosis inception rates with different missing delay estimates and confidence intervals of the inception rates based on the median of the CDD (as a ratio of inception rates based on the median of the CDD) for smokers and durations 0 & 1. 142
7.27 Graphs of relative diagnosis inception rates with different missing delay estimates and confidence intervals of the inception rates based on the median of the CDD (as a ratio of inception rates based on the median of the CDD) for smokers and durations 2 & 3. 143
7.28 Graphs of relative diagnosis inception rates with different missing delay estimates and confidence intervals of the inception rates based on the median of the CDD (as a ratio of inception rates based on the median of the CDD) for smokers and durations 4 & 5+. 144

8.1 Model selection for CABG. 148
8.2 Graphs of diagnosis inception rates for CABG for males, non–smokers (MNS) and smokers (MS). 150

8.3 Graphs of diagnosis inception rates for CABG for females, non–smokers (FNS) and smokers (FS). 151
8.4 Model selection for cancer. 152
8.5 Graphs of diagnosis inception rates for cancer for males, non–smokers (MNS) together with CMI rates. 155
8.6 Graphs of diagnosis inception rates for cancer for males, non–smokers (MNS) and smokers (MS). 156
8.7 Graphs of diagnosis inception rates for cancer for females, non–smokers (FNS) and smokers (FS). 157
8.8 Graphs of diagnosis inception rates for cancer for males, non–smokers vs smokers. 158
8.9 Model selection for death. 161
8.10 Graphs of diagnosis inception rates for death for males, non–smokers (MNS) and smokers (MS). 162
8.11 Graphs of diagnosis inception rates for death for females, non–smokers (FNS) and smokers (FS). 163
8.12 Graphs of diagnosis inception rates for death for males, non–smokers vs. smokers. 164
8.13 Model selection for heart attack. 165
8.14 Graphs of diagnosis inception rates for heart attack for males, non–smokers (MNS) and smokers (MS). 167
8.15 Graphs of diagnosis inception rates for heart attack for females, non–smokers (FNS) and smokers (FS). 168
8.16 Model selection for kidney failure. 169
8.17 Graphs of diagnosis inception rates for kidney failure for males (M) and females (F). 171
8.18 Model selection for major organ transplant. 172
8.19 Graph of diagnosis inception rates for major organ transplant for all population. 173
8.20 Model selection for multiple sclerosis. 174
8.21 Graphs of diagnosis inception rates for multiple sclerosis for males, nonsmokers (MNS) and smokers (MS). 176

8.22 Graphs of diagnosis inception rates for multiple sclerosis for females, nonsmokers (FNS) and smokers (FS). 177
8.23 Model selection for other causes. 178
8.24 Graphs of diagnosis inception rates for other diseases for males, full accelerated (MFA) and stand alone policies (MSA). 180
8.25 Graphs of diagnosis inception rates for other diseases for females, full accelerated (FFA) and stand alone policies (FSA). 181
8.26 Model selection for stroke. 182
8.27 Graphs of diagnosis inception rates for stroke for males, nonsmokers (MNS) and smokers (MS). 184
8.28 Graphs of diagnosis inception rates for stroke for females, nonsmokers (FNS) and smokers (FS). 185
8.29 Graphs of diagnosis inception rates for stroke for males, nonsmokers vs smokers. 186
8.30 Model selection for TPD. 187
8.31 Graphs of diagnosis inception rates for TPD for policy durations 0 (PD0) and 1 (PD1). 189
8.32 Graphs of diagnosis inception rates for TPD for policy durations 2 (PD2) and 3 (PD3). 190
8.33 Graphs of diagnosis inception rates for TPD for policy durations 4 (PD4) and 5+ (PD5+). 191
8.34 Contribution of individual causes for males, full accelerated policies, non-smokers, year 2003, policy durations 3 and Office1. 195
8.35 Comparison of all-cause rates and summation of cause-specific rates for males, full accelerated policies, non-smokers, year 2003, policy duration 3 and Office1. 196
8.36 Contribution of individual causes for females, full accelerated policies, non-smokers, year 2003, policy durations 3 and Office1. 197
8.37 Comparison of all-cause rates and summation of cause-specific rates for females, full accelerated policies, non-smokers, year 2003, policy duration 3 and Office1. 198

8.38 Contribution of individual causes for males, full accelerated policies, smokers, year 2003, policy durations 3 and Office1. 199
8.39 Comparison of all-cause rates and summation of cause-specific rates for males, full accelerated policies, smokers, year 2003, policy durations 3 and Office1. 200
8.40 Contribution of individual causes for females, full accelerated policies, smokers, year 2003, policy durations 3 and Office1. 201
8.41 Comparison of all-cause rates and summation of cause-specific rates for females, full accelerated policies, smokers, year 2003, policy duration 3 and Office1. 202

List of Abbreviations

ABI	The Association of British Insurers
CABG	Coronary artery by-pass graft
CDD	Claim delay distribution
CII	Critical Illness Insurance
CMI	The Continuous Mortality Investigation
FA	Full Accelerated
GL-type	Generalised Linear type
GVS	Gibbs Variable Selection
HA	Heart Attack
IGa	Inverse Gamma
KF	Kidney Failure
MOT	Major Organ Transplant
MS	Multiple Sclerosis
SA	Stand Alone
TPD	Total and Permanent Disability
WP	Working Paper

Chapter 1

Introduction

1.1 Critical Illness Insurance

Critical Illness Insurance (CII) is a type of long term insurance that provides a lump sum on the diagnosis of one of a specified list of critical illnesses within the policy conditions. CII first came to the scene in South Africa early in the 1980s under the name of Dread Disease Insurance. However, before this, in the USA, Japan and Israel some life insurance policies were extended to cover cancer (Dash and Grimshaw, 1993).

CII has been very popular in the UK. Although CII policies have been issued since the 1980s in the UK, the number of policies increased dramatically in the early 1990s. From then on, it continued growing. While there were 100000 new sales in 1990, this figure increased almost 7 times to 700000 new CI policies sold in 1998 (Dinani *et al.*, 2000). More recent sales figures reveal that more than one million new policies were issued in 2002 (CMI WP 50, 2011). The data we are using throughout the thesis cover the period between 1999 and 2005 and the inforce figures also indicate a positive growth during the period, however with a smaller rate.

There are some particular aspects of CII which make this type of policy attractive, e.g. the lump sum benefit is payable on diagnosis regardless of the duration of the illness. This can make CII more popular than permanent health insurance. Moreover, a single person who has no dependants would probably prefer CII instead of buying

conventional life assurance in order to meet the high expenses of serious illness (Dash and Grimshaw, 1993). There is no restriction on how to spend the CII benefit. Most of the CII policies in the UK are linked to mortgages as this is a considerable financial commitment and diagnosis with a critical illness could affect the individual's ability to repay the mortgage.

In the UK, there are two types of CI policy: Full Accelerated (FA), which covers both critical illness and death, and Stand Alone (SA), which covers only critical illness. Most of the policies are accelerated policies and they are attached to life insurance, term insurance or endowments (see WP 50 (2011)). Typically, regular premiums are payable throughout the term while the policy is in force.

CII coverage includes, but is not limited to, cancer, heart attack, stroke, coronary artery by-pass graft (CABG), kidney failure, major organ transplant (MOT) and multiple sclerosis (MS). Most policies include total and permanent disability (TPD) for completeness, essentially to cover disability arising from other reasons which might not be covered explicitly by other causes. These 8 diseases and death form more than 90% of the claims. However the CII market in the UK is very competitive so that insurance companies may increase the number of illnesses covered in an attempt to increase their market share.

One feature of CII is that it has mostly clear and understandable definitions of what constitutes a claim. In the UK, the Association of British Insurers (ABI) has defined the illnesses covered by CII. These definitions are presented in ABI (2006) and for most of the illnesses are accepted as a standard guide by the insurance companies. However, unlike most of the other causes covered by CII, TPD has a vague definition (e.g 'own or any occupation' statement) and this leads to some problems such as rejection of many claims as valid or long waiting periods until the settlement of a claim. To reduce the ambiguity in these claims and raise the understanding of consumers, recently, the ABI set some specific model definitions using their past experience. These definitions are explained in detail in ABI (2010). However note that the data used in this thesis do not cover these recent definitions.

In general, CII is subject to long delays between the dates of diagnosis and settlement, which in some cases can be measured in years rather than in weeks or months, and

this leads to some problems such as incurred but not settled (IBNS) claims. This is an important problem, as not allowing for this in the modelling will distort the results by exposure year (because of the removal of claims settled in the exposure year but diagnosed in earlier years and addition of claims diagnosed in the exposure year but settled in later years). This issue is also raised by the Continuous Mortality Investigation (CMI) (see WP 14 (2005), WP 28 (2007) and WP 33 (2008)).

The CMI is a research body organised by the Actuarial Profession in the UK. It collects data from contributing life insurance offices and carries out research on all main life insurance branches including critical illness. It publishes the research in reports and working papers. The data we used in this thesis were supplied by the CMI and they include claims settled between 1999 and 2005. We were provided with a large set of claims data and the associated in force data collected from contributing life insurance offices in the UK. This data set represents about one half to one third of the CII policies sold in the UK for that period (WP 50, 2011). The data will be explained in more detail in Chapter 2.

1.2 Overview of past results

Our ultimate aim in this thesis is modelling diagnosis inception rates for CII using appropriate statistical methodology. The cost of CII based on the UK data was modelled by Dash and Grimshaw (1993) and some incidence rates for heart attack, stroke, and cancer were produced for full accelerated policies. A base table called IC94 was produced by the Working Party of the Society of Actuaries in Ireland (1994). In that table the rates were mostly derived from population data in the UK, however the data were adjusted for Ireland and insured lives. The rates were produced for different ages and sex but not for smoker status. The CI market in the UK between 1991 and 1998 is examined by Dinani *et al.* (2000) and a base table, CIBT93, was published, mostly from English population data between 1993-1994. The data were obtained from the Office of National Statistics, Hospital Episodes Statistics and Morbidity Statistics from General Practice. Incidence rates were calculated for the core diseases and they were applicable to both full accelerated and stand alone policies for different sexes and ages. The incidence rates for the full accelerated policies were calculated

using Dash and Grimshaw's (1993) model for the UK population. This base table was updated in 2006 by the CI Trends Research Group with the 1999-2002 experience using the same data source and the CIBT02 base table was produced. This base table was also for the UK population. As for CIBT93, the rates were age and sex specific but they did not differ by smoker status. CIIT00 which was produced by GenRe in 2007 was an attempt to construct a table using insured lives data. Basically the population incidence rates from CIBT02 base table were rescaled using the 1999-2002 insured experience of the CMI because they mention that the size of the insured data set was not sufficient to construct a new table. Different incidence rates for smokers and non–smokers were provided in this table. Two recent studies by the CMI give CI inception rates in the UK, namely WP 43 (2010) and WP 50 (2011). These rates are based on insured data from the UK. These two studies will be discussed in more detail here, as a more comprehensive point of view is provided compared to the other studies. Nevertheless, none of the incidence rates for CII produced so far depends on a statistical model.

In this thesis we first model the delay between diagnosis and settlement of claims as this delay can be very long and subject to uncertainty in CII. These long delays lead to problems of incurred but not reported (IBNR), incurred but not settled (IBNS) and reported but not settled (RBNS) claims. This issue is also mentioned by the CMI (in WP 14 (2005), WP 28 (2007) and WP 33 (2008)) and the necessity of an adjustment factor to avoid the understatement of the experience is stressed. This factor is expected to be dependent at least on office as well as other characteristics of the claim, such as cause of claim, smoker status, policy duration. Here, we consider a generalised model setting with a three-parameter Burr distribution to model the claim delays and estimate parameters under both a classical approach and a Bayesian approach using Markov Chain Monte Carlo (MCMC). The three-parameter Burr distribution has many applications in actuarial science due to its high flexibility in modelling heavy tails. Beirlant *et al.* (1998) extended the Burr distribution to a regression model by allowing either one of its shape parameters or the scale parameter to vary with the covariates. In their paper, they estimated the parameters by using a maximum likelihood approach. Following this, Beirlant and Guillou (2001) considered extreme value methods for Pareto-type distributions under censoring, and discussed maximum likeli-

hood estimates of the extreme value index of the Burr distribution, which is defined in terms of shape and scale parameters. Explanatory variables are also regressed on this index by Beirlant and Goegebeur (2003) and maximum likelihood estimates are given. Regression-type models with heavy-tail errors are considered in a Bayesian context by Antonio and Beirlant (2008) and Frees and Valdez (2008). MCMC methodology for actuarial related problems using the Burr distribution is discussed by Scollnik (2001).

In order to obtain the most suitable model for describing and predicting the delay between diagnosis and settlement of claims, we also investigate Bayesian variable selection among the available claim-related factors in this study. Different prior distribution settings for the model parameters are considered employing methodology introduced by Dellaportas *et al.* (2002), Ntzoufras (2002, 2009) and Ntzoufras *et al.* (2003).

One of the biggest difficulties the CMI encountered in providing inception rates is that of non-recorded dates of diagnosis. In WP 14 (2005) and WP 28 (2007) the investigation period is 1999-2002 and only 56% of the claims have the date of diagnosis recorded. The number of recorded dates of diagnosis increased to 70% in 2003 and reached 75% in 2004 according to WP 33 (2008). Eventually in WP 50 (2011) it is mentioned that, for the period between 2003 and 2006, approximately 80% of the claims have date of diagnosis available.

Inception rates were smoothed by the CMI for CII and the results were presented in the CMI's WP 43 (2010) and WP 50 (2011). These two studies are the most relevant ones to our work. The CMI's methodology to produce diagnosis inception rates is explained in WP 43 (2010) where the 1999-2004 experience is used. In both of these working papers, an initial set of rates is adjusted using the CIBT02 base table to obtain the inception rate. Nonetheless, not using adequate mathematical modelling techniques creates difficulties. One of the problems with this method is that it is applied to subsets of the data. When the volume of the data is not enough, the rates for that subgroup can not be derived. Therefore the analyses of the rates in these working papers are limited to full accelerated policies only. The rates can only be produced for some combinations of sex, smoker status and policy duration. The number of combinations is increased in WP 50 (2011) which uses the 2003-2006

experience and it is mentioned that the data are more stable in this period. In WP 43 (2010), cause-specific rates (only for cancer, heart attack, death, stroke, CABG and TPD) are presented for male, non–smokers, curtate policy durations 0 (duration of policy is less than a year) to 5+ (duration of policy is more than 5 years). Since the method used there is highly data-intensive, the rates are not reliable outside the age range 25-65 for the all-cause rates while it is restricted to 30-60 for cause-specific rates. Also, as the data have a significant amount of missing dates of diagnosis, the claims are matched to exposure, according to their year of settlement (not diagnosis). In addition to these disadvantages, another, and probably more serious, problem with this non-statistical method is that no confidence intervals can be given for the rates because it is not possible to calculate a standard deviation with this method.

In graduating claim diagnosis rates, we use a similar model to that described in Forfar *et al.* (1988). A good discussion of smoothing rates can be found in this paper where the 1979-1982 mortality experience is used for the UK.

1.3 Stochastic modelling for inception rates

Regardless of the insurance type, insurance companies need to assess the diverse range of risks they face as this will affect their evaluation of future cash flows, premium calculations and policy values in order to be solvent and meet the capital requirements. Morbidity is one of the risks for health related policy products that insurance companies have to manage effectively.

The occurrence and amount of liabilities are the inherent uncertainties about morbidity. The risk can be reduced if claim inception rates are modelled suitably and therefore it is important to develop a stochastic model of morbidity risks. Some factors such as age, sex or smoker status are most likely to have an effect on morbidity rates. Also, consequences of different benefit amounts, years of diagnosis or policy durations can be investigated. Depending on the insurance type, there might be other factors as well. For example, for CII, policy type (full accelerated/stand alone) could be an important factor in terms of claim inception rates. Apart from these, different offices might experience different morbidity rates due to their underwriting proce-

dures, market coverage (different target of socio-economic groups) or interpretations of cause of claims. On the other hand, external factors such as medical advances (new treatments, new diagnostic techniques that may cause illness definitions to change) or changes in lifestyles (social, behavioural) are difficult to represent mathematically.

In this thesis we will apply statistical methodology to estimate the claim inception rates for CII by date of diagnosis and by cause. This will be the first statistical model for the diagnosis inception rates for CII.

We consider 8 causes of claims and death. All other causes are collected under a category comprising 'Other' causes which corresponds to 6.6% of the data (details about grouping illnesses are discussed in Chapter 2). Considering all of the CI causes we have covered, we have 11 possible states for considered policyholders, including the healthy state. A policyholder is supposed to be healthy at the time of the commencement of the policy and he/she stays in this state until at some future time he/she transits to one of the 10 possible exit states. These states and the possible transitions are represented in Figure 1.1.

As mentioned in Section 1.1, a CII ceases on the payment of the lump sum. Therefore all CI cause states are absorbing states. Since transitions between these states are not possible in both directions, the model, can be considered as a multiple decrement model (see Waters (1984)).

In Figure 1.1, the transition intensities are denoted by $\lambda^c(x;\boldsymbol{\theta})$ where $c = \{$CABG, Cancer, Heart Attack, Kidney Failure, MOT, Multiple Sclerosis, Stroke, TPD, Other, Death$\}$ and they are allowed to depend on the policyholder's attained age (x) as well as other characteristics (e.g. sex, smoker status, policy type, office). These characteristics are denoted by a vector, $\boldsymbol{\theta}$. Here, we emphasize that different incidence rates may depend on different subsets of $\boldsymbol{\theta}$ since covariates are selected to provide the most appropriate model for each transition. This will be discussed later in Chapter 8 when we model the incidence rates for individual causes.

In Chapter 7 we will model the incidence rates for 'all-causes', $\lambda(x;\boldsymbol{\theta})$, where

$$\lambda(x;\boldsymbol{\theta}) \approx \sum_c \lambda^c(x;\boldsymbol{\theta}).$$

We do not use a strict equal sign in this equation, as the equivalence is relative to different exposures, as will be discussed in Chapter 8. In our model, maximum likelihood estimation of the intensity under the Markov model provides the same estimates as when the observed claim numbers, $N(x;\boldsymbol{\theta})$, are assumed to follow a Poisson$(\lambda(x;\boldsymbol{\theta})E^*(x;\boldsymbol{\theta}))$ distribution for known exposure (here $E^*(x;\boldsymbol{\theta})$ can be regarded as 'adjusted exposure' where the cdf of the appropriate claim delay distribution (CDD) is used as an adjustment factor). Detailed explanation and the estimation procedure can be found in Macdonald (1996a, b, c). Therefore in Chapters 7 and 8 we will assume that the observed number of claims has a Poisson distribution. This will facilitate the use of a generalised linear model (GLM) for smoothing the inception rates.

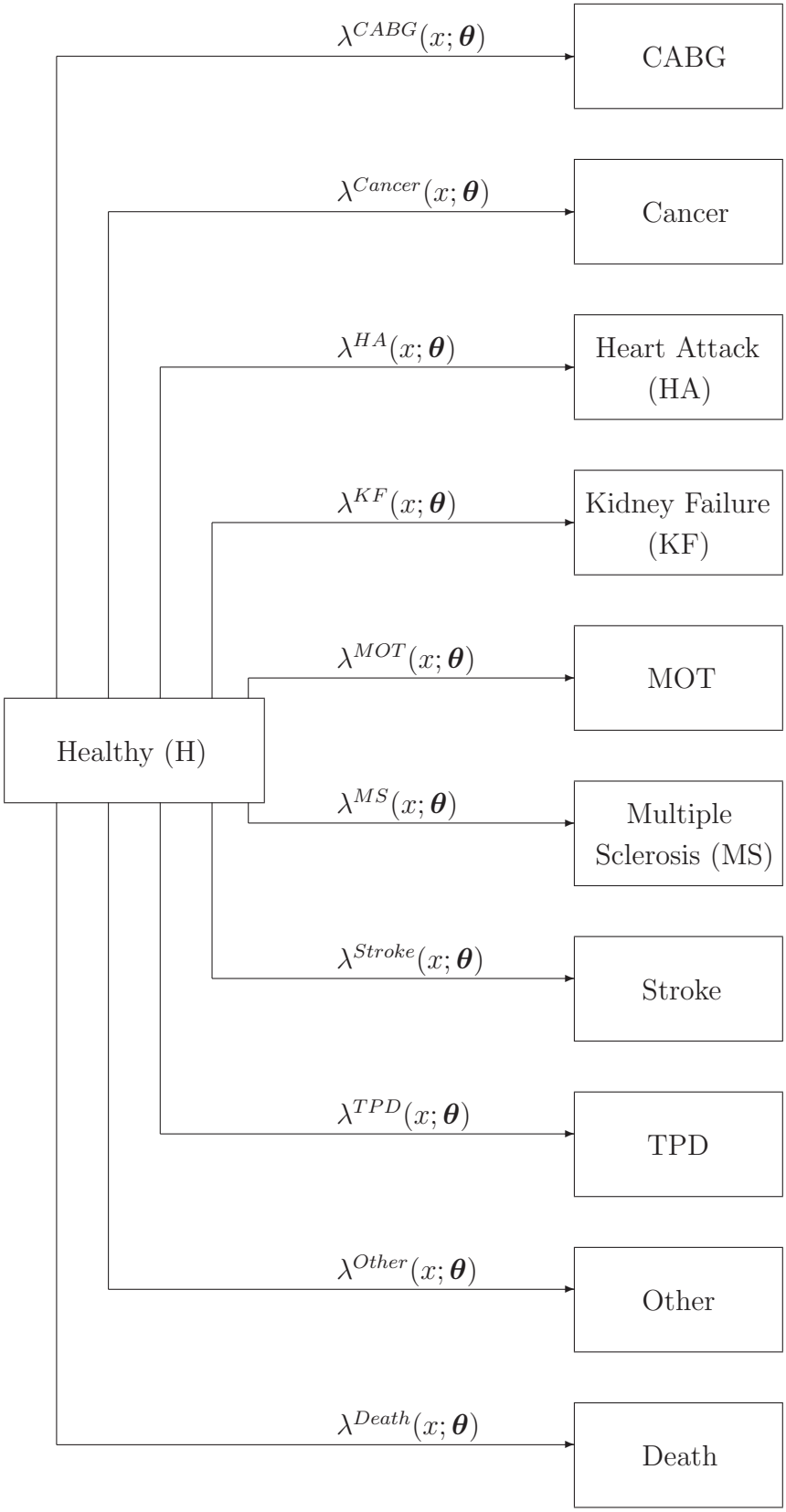


Figure 1.1: A multi-state model for specific causes for CII.

1.4 Outline of the thesis

The layout of the thesis is as follows:

The data we used are described in Chapter 2. The information available for each observation is presented and the grouping of the CI causes is described. This grouping is based on core illnesses defined by the ABI and on CMI experience.

In Chapter 3 the modelling of the claim delay distribution (CDD) without considering business growth and missing observations is presented. As the Burr is not a member of the exponential family of distributions, a generalised-linear-type model (GL-type) is fitted, where claim-related factors possibly affecting the delay are linked to the mean of the distribution. The method we use to estimate the CDD gives us an opportunity to estimate the claim delays for various characteristics in a single model and calculate the probability of settling a claim using an appropriate CDD.

When modelling the claim delays in Chapter 4, changes in the volume of CI business are taken into consideration for each office; the growth rates differ significantly between individual offices.

The most suitable CDD is chosen in Chapter 5 after variable selection using both Bayesian methodology under different prior distribution setups and frequency-based model selection techniques. First, claim specific covariates are selected for the models without considering the growth rates, and the selection is then repeated for the model which includes growth rates.

Although the number of submitted dates of diagnosis is increasing over time, the number of missing cases is still significant considering that date of diagnosis is accepted as the 'date of claim'. Excluding these claims from the analyses might lead to biased estimates, especially if these missing dates are not random (e.g. depending on office or cause). In this thesis we use a Bayesian approach as this incorporates missing or unobserved information in a natural manner. Under a Bayesian setting missing data are treated as additional parameters that are estimated in the analysis through imputation of their values in a MCMC scheme in Chapter 6. The missing values are estimated using the median of the most suitable CDD obtained in this chapter and this complete data set is used when the diagnosis inception rates are modelled.

The estimation of diagnosis inception rates for all-causes (combined) is presented in Chapter 7. An appropriate CDD is used to adjust the exposures in order to eliminate the effects of long delays between dates of diagnosis and settlement of claims (e.g. IBNS). The number of claims is assumed to have a Poisson distribution and diagnosis inception rates are smoothed using a similar model to that described in Forfar *et al.* (1988). Diagnosis inception rates are smoothed using three different subsets of covariates. In the first model, we used the best model after variable selection. To compare our rates with CMI rates (in WP 43 (2010)) we also used the best model which includes the CMI variables and the model which only uses the CMI variables. Here, CMI variables refer to policy type, sex, smoker status and policy duration variables. Also, in this chapter, the effect of using the median of the CDD obtained in Chapter 6 to estimate the missing dates on the diagnosis inception rates is investigated by a sensitivity analysis.

Diagnosis inception rates are modelled for specific causes and results are compared with all-cause rates in Chapter 8.

Finally, our conclusions together with some discussions and ideas for further research are presented in Chapter 9.

Chapter 2

Data

Our aim is to develop models for critical illness insurance for providing diagnosis inception rates by determining the effects of various risk factors on diseases and death. To do this, we use the claims data and corresponding inforce data supplied by the CMI in the UK. The CMI collects data from contributing offices on critical illness business on a calendar year basis. Since CI policies can have long periods between the dates of diagnosis and settlement, it is not practical for the CMI to wait until all the diagnosed claims in a particular year have been settled. Hence, the CMI asks for the submission of the settled claims during the year only (see e.g. CMI WP 33 (2008)). The data with which we are provided relate to claims inforce and/or settled in the years 1999 - 2005.

Each observation in the data set has various characteristics. These can be listed as

(a) Sex: female, male are coded as F and M, respectively

(b) Smoker status: non-smoker, smoker, undifferentiated are coded N, S and U, respectively

(c) Record year: taking values from 1999 to 2005

(d) Benefit type: full accelerated, stand alone are coded as F and S, respectively

(e) Office number: coded anonymously from letter A to letter S. If the office has changed its portfolio of policies during its contribution period, this is shown by a suffix (e.g. A becomes A2 and then A3)

(f) Territory: the UK, Eire are coded as 1 and 2 (all of the claims have code 1), respectively

(g) Policy type: joint or single life are coded as J and S, respectively

(h) Sales channel: bancassurer, direct sales, IFA, other, unknown are coded as B, D, I, O and U, respectively

(i) Benefit amount: in GBP

(j) Submission group: coded from 1 to 15 to show the contribution period of an office (see Table 2.1 for details)

(k) Date of birth: in DDMMYYYY format

(l) Date of commencement (of the policy): in DDMMYYYY format

For the claims data, we are also provided with

(m) Type of claim: death, critical illness are coded as D and C, respectively

(n) Cause of critical illness claim: there are 54 different critical illnesses and death giving cause of claim in the data set.

Moreover for each claim, contributing offices are asked to provide 4 dates which are

(o) Date of diagnosis: in DDMMYYYY format

(p) Date of notification: in DDMMYYYY format

(q) Date of claim admission: in DDMMYYYY format

(r) Date of settlement: in DDMMYYYY format.

However these dates are not always completed by the offices and sometimes they show inconsistencies (e.g. date of notification is earlier than date of diagnosis).

Table 2.1: Details of the submission groups.

Group	Years of submission
1	Data submitted consistently throughout 1999-2005
2	Data submitted consistently throughout 1999-2002
3	Data submitted consistently throughout 2003-2005
4	Data submitted consistently throughout 2003-2004
5	Data submitted for 2005 only
6	Data submitted consistently throughout 1999-2003
7	Data submitted for 2004 only
8	Data submitted consistently throughout 1999-2001
9	Data submitted consistently throughout 2000-2005
10	Data submitted consistently throughout 2000-2004
11	Data submitted for 1999 only
12	Data submitted consistently throughout 2000-2002
13	Data submitted for 2002 only
14	Data submitted consistently throughout 2000-2003
15	Data submitted consistently throughout 2004-2005

These submission groups enable us to determine details of the policies inforce at the start and at the end of each of the seven years of the investigation period. This information is needed to calculate exposure - see Chapter 7.

Another point to mention is that we have not used 'Sales Channel' in our analyses since it is highly related to 'Office'. Of the 13 offices we have considered in our analyses, 6 offices use only one sales channel among four known channels. The numbers of sales channels used by the offices are given in Table 2.2. One of the offices uses only the 'Unknown' sales channel, thus it is not shown in the table. Since offices use specific sales channels, office and sales channel are highly related and therefore one of these two variables is unnecessary in the analyses.

Table 2.2: Number of offices and number of sales channels.

	No. sales channels			
	1	2	3	4
No. offices	6	3	1	2

The claims file contains records of 27244 claims settled in the seven year period from 1999 to 2005. Of these claims, a total of 8117 (29.8%) were omitted from our analysis

of the claim delay distribution for one of the following reasons:

(i) One of the offices provided both dates of diagnosis and settlement for only 235 claims out of 4426 and all of them are death claims. This caused many computational problems when including the missing values as we do not have sufficient information for the other causes for this office. (4426)

(ii) Both date of diagnosis and date of settlement are missing (some offices have never provided both dates of diagnosis and settlement). (3585)

(iii) Smoker status is not recorded. (61)

(iv) Date of commencement is the same as date of diagnosis. (42)

(v) Date of commencement is the same as date of birth. (1)

(vi) Date of commencement is the same as date of notification. (1)

(vii) Date of commencement is the same as date of settlement. (1)

As a result we ended up with 19127 claims. In the data, 9.2% of these claims have no date of diagnosis which we use as date of claim. We need the date of claim to calculate the duration of the policy and policyholder's age at the time of claim. By defining the date of claim, claims will be assigned to a particular year's experience. If the aim is to produce claim diagnosis rates, the date of diagnosis seems suitably to represent the claim date, as the claim will have been incurred in the insurance period but not necessarily reported or settled in the same year. In this respect, it reflects the true cost to the insurer (CMI WP 14, 2005; CMI WP 33, 2008).

On the other hand, claims are submitted to the CMI according to their settlement year and 7.9% of the data have no date of settlement. Out of 19127 claims settled in the period 1999 - 2005, only 15860 have both dates of diagnosis and settlement. This corresponds to 82.9% of the whole claims data set. The observed intervals between the four dates relate to each claim and the missing figures for these delays are summarised in Table 2.3.

To summarise the data, the number of claims and percentages by various factors are given in Table 2.4. The identity of each contributing office is confidential so this characteristic was recorded as a code. Our data originated from 13 different offices, not

Table 2.3: Average observed delays between dates of diagnosis, notification, admission and settlement (in days).

	Diagnosis to Notification	Notification to Admission	Admission to Settlement	Diagnosis to Settlement
Mean delay	93	80	18	185
No. observations	15585	9190	9752	15860
% of observations having both dates	81%	48%	51%	83%

all of which contributed data in each of the seven years. In terms of the percentage of the total claims, the largest contribution was 27.5%, the smallest 0.1% and the median was 3.0%.

Table 2.4: Number of claims and percentages by various factors.

Benefit Type		**Critical Illnesses**	
Full Accelerated	16875 (88.2%)	CABG	393 (2.1%)
Stand Alone	2252 (11.8%)	Cancer	9381 (49.0%)
		Deaths	3371 (17.6%)
Joint/Single Life		Heart Attack	2220 (11.6%)
Joint Life	9743 (50.9%)	Kidney Failure	110 (0.6%)
Single Life	9384 (49.1%)	Major Organ Transplant	36 (0.2%)
		Multiple Sclerosis	825 (4.3%)
Gender		Other	1265 (6.6%)
Female	8173 (42.7%)	Stroke	1027 (5.4%)
Male	10954 (57.3%)	Total and Permanent Disability	499 (2.6%)
Smoker Status		**Type of Claim**	
Non-Smoker	14129 (73.9%)	Claim (Critical Illness)	15756 (82.4%)
Smoker	4998 (26.1%)	Death	3371 (17.6%)

Most of the policies (88%) are covering death as well as critical illness, i.e. full accelerated policies. The greater proportion of claimants are non-smokers; the non-smoker/smoker split (75%/25%) is approximately the same as in the general UK population. We have a good representation for females and males. Joint life and single life policies are almost equally split.

CII policies cover a great number of illnesses, but eight illnesses – cancer, heart attack, stroke, CABG, MOT, kidney failure, MS and TPD – and death form 93.4% of the total cases. All other causes are grouped in the 'Other' claim cause category which includes 6.6% of the data. In total, 10 causes of claims are taken into consideration in our analyses. These diseases and relevant groupings are explained in detail later in this section.

Grouping the critical illness causes

There are 55 different cause codes in the claims data, most of which are types of cancer. Among the specified cancer types, the biggest group is female breast cancer with 1838 claims. Since we have a significant amount of data for this cancer type we wanted to analyse it as a separate cause. However cancer claims include 'site not specified' which, in fact, includes the biggest number of claims with 4363 claims out of a total 19127 claims (and corresponds to 22.3% of all claims and 45.4% of all cancer claims). We are told by the CMI that 'site not specified' group includes female breast cancer claims as well as other cancers. This means that analysing cancer types individually will underestimate the true rates. Therefore we do not subdivide cancer into categories and analyse cancer as a single cause. This issue is also mentioned by the CMI in WP 43 (2010, page 38).

The 55 causes of claim and the groups used in the analyses are given in Table 2.5. In the table, the number of claims for each cause is also presented. We reduce the number of causes for analysis by grouping some of the less frequent causes mostly depending on the ABI definition of 'core' conditions (ABI, 2005) and on the grouping which the CMI uses in its working papers (this is explained in CMI WP 14 (2005, Appendix C)).

Cancer, heart attack and stroke are the major causes of claim as they are always covered under a critical illness policy. According to the ABI (2005), other core conditions are CABG, kidney failure, MOT and MS. Apart from these diseases we treat TPD as a separate group following CMI practice. All the other causes merged into the 'Other' cause group (unlike the CMI's WP 43 (2010), we group terminal illness and unknown in other causes. Later in WP 50 (2011), unknown claims are combined with the other

causes by the CMI). These 9 categories together with death form our cause variable throughout the thesis.

Table 2.5: Grouping cause of claim.

	CMI Cause	Coded as	No. Claims
1	Coronary Artery Bypass Graft (CABG)	CABG	393
2	Cancer - site not specified	Cancer	4363
3	Hodgkin's disease	Cancer	60
4	Leukaemia	Cancer	201
5	Malignant melanoma of skin	Cancer	285
6	Malignant Neoplasm (MN) of oesophagus	Cancer	29
7	MN - multiple sites	Cancer	19
8	MN of bladder	Cancer	56
9	MN of bone and articular cartilage	Cancer	12
10	MN of bone, connective tissue, skin and breast - unspec.	Cancer	16
11	MN of brain	Cancer	128
12	MN of colon	Cancer	286
13	MN of digestive organs and peritoneum - unspec.	Cancer	44
14	MN of female breast	Cancer	1838
15	MN of genitourinary organs - unspec.	Cancer	78
16	MN of kidney and other urinary organs	Cancer	111
17	MN of larynx	Cancer	40
18	MN of lip, oral cavity and pharynx	Cancer	37
19	MN of liver	Cancer	46
20	MN of lymphatic and haematopoietic tissue	Cancer	279
21	MN of other sites	Cancer	302
22	MN of ovary and uterine adnexa	Cancer	229
23	MN of pancreas	Cancer	63
24	MN of prostate	Cancer	199
25	MN of rectum, rectosigmoid junction and anus	Cancer	8
26	MN of respiratory and intrathoracic organs - unspec.	Cancer	11
27	MN of small intestine including duodenum	Cancer	18
28	MN of stomach	Cancer	64
29	MN of testis	Cancer	361
30	MN of trachea, bronchus and lung	Cancer	228
31	Myeloid leukaemia	Cancer	8
32	Other MN of skin	Cancer	22
33	Deaths	Death	3371
34	Heart Attack (HA)	HA	2220
35	Kidney Failure (KF)	KF	110
36	Major Organ Transplant (MOT)	MOT	36
37	Multiple Sclerosis (MS)	MS	825
38	Alzheimers Disease	Other	8
39	Angioplasty	Other	131
40	Aorta Graft Surgery	Other	15
41	Benign Brain Tumour	Other	278
42	Blindness	Other	7
43	Coma	Other	74
44	Deafness	Other	1
45	Heart Valve Replacement / Repair	Other	151
46	Loss of limbs	Other	5
47	Motor Neurone Disease	Other	42
48	Other	Other	208
49	Paralysis / Paraplegia	Other	47
50	Parkinsons Disease	Other	66
51	Terminal Illness	Other	6
52	Third Degree Burns	Other	5
53	Unknown	Other	161
54	Stroke	Stroke	1027
55	Total Permanent Disability (TPD)	TPD	499

Chapter 3

Modelling CDD I: Without considering the missing values and growth rates

3.1 Introduction

The aim of this chapter is to model the delay between diagnosis and settlement dates. Both classical methods and Bayesian methods are employed. The classical method is based on maximum likelihood estimation. The coefficients are obtained by maximising the likelihood with a Newton-Raphson iterative approach using R software (R Development Core Team, 2009). In Bayesian analyses we consider the posterior distribution, which is given in terms of the likelihood function and the prior distribution. However, in some cases these posterior distributions might be too complex to be derived analytically. In these cases, MCMC methods can be used to integrate over the high dimensional posterior distribution and make inferences about model parameters (Gilks *et al.*, 1996). Bayesian methodology is quite natural for interpreting statistical conclusions and its flexibility allows complex data sets to be managed (Gelman *et al.*, 2000). Moreover, MCMC methodology allows us to apply Bayesian analysis to the complex data sets considered here and also to include missing observations in later chapters. Bayesian methods and MCMC techniques are used with the help of the WinBUGS software package (Spiegelhalter *et al.*, 2003). In this chapter, we only

consider claims where both dates of diagnosis and settlement are known. There are 15860 such cases out of 19127 which corresponds approximately to 83% of the data. Later, in Chapter 6, we also include unobserved dates in the analysis.

Some features of the observed delay between the dates of diagnosis to settlement are given in Section 3.2. In Section 3.3, we present the results assuming the contributing offices do not grow within successive years. In Section 3.4, prediction results are given for some hypothetical scenarios.

3.2 The observed delay

Before we start modelling we would like to give some features of the delay between the dates of diagnosis and settlement. The mean delay between these two dates is 185 days whereas the median is 111 days and the maximum delay is 3980 days. Figure 3.1 shows the distribution of the duration of the delay for these claims, measured in days. Although in theory it is not likely to be diagnosed and settled in the same day, some claims have zero days between these two dates. Since we are told by the CMI that these claims are genuine claims and there are no known mistakes, we kept these 39 claims in the analyses by adding 0.5 days to delay due to the calculation problems (logarithmic transforms and distributions defined only for strictly positive values). Most of these claims (31) are death claims with 4 cancer, 3 TPD and 1 other causes claims.

Without any modelling, the box plot of the delay by cause on a logarithmic scale is shown in Figure 3.2. The interquartile range of TPD is the widest. The median of the delay of death claims is 78 days and it is smaller than the medians of other causes. The minimum delay of MOT is the longest one with 24 days.

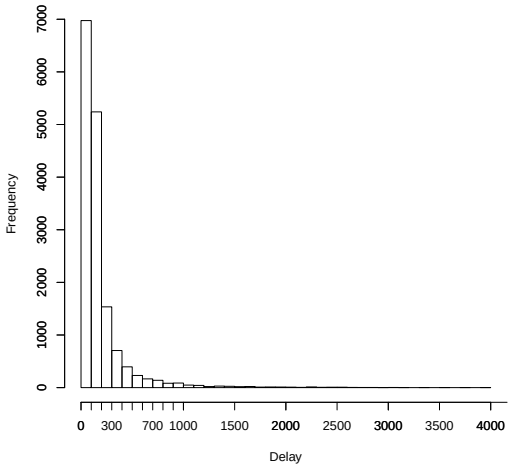


Figure 3.1: Histogram of claim settlement delay (in days).

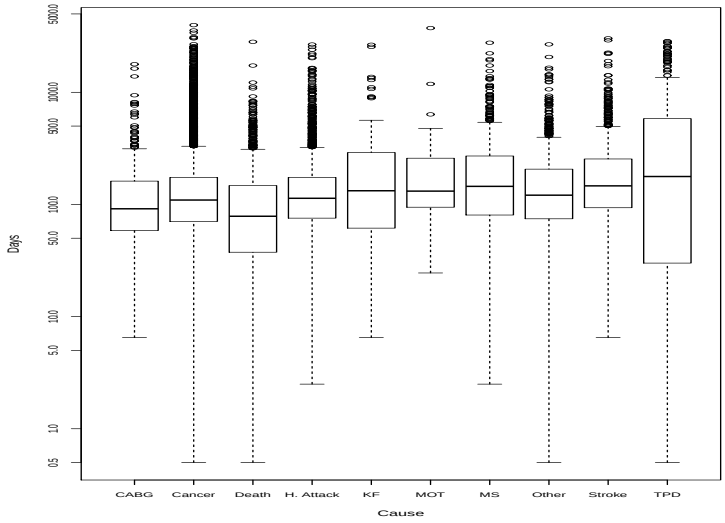


Figure 3.2: Box plots of observed delay by cause (in days).

3.3 Modelling under the assumption of no growth within offices

Without subdividing the data set by different risk factors, parametric models are fitted to 15860 claims where we have dates of both diagnosis and settlement. Considering the shape of the histogram in Figure 3.1, the data are positively skewed which is common for waiting time data, especially in insurance. Commonly, standard 2-parameter distributions such as the lognormal, Pareto and Gamma are employed in insurance data problems, partly because of their straightforward use. Here the distribution of the observed data suggests that a 3-parameter distribution may be more suitable due to its flexibility in modelling heavy-tails. We note that the Pareto distribution is a special case of the Burr distribution when one of the shape parameters is equal to 1.

We use a 3-parameter Burr model and compare our main findings with results obtained using the more commonly used lognormal distribution. We treat all unknown coefficients and parameters appearing in the model as random quantities and we assign to them prior distributions to complete the Bayesian setting.

The Burr model

The delay for claim i is denoted by a random variable D_i and under the assumed Burr(α, τ, λ) model, its probability density function is given by

$$f(d_i; \alpha, \tau, \lambda) = \frac{\alpha\tau\lambda^{\alpha} d_i^{\tau-1}}{(\lambda + d_i^{\tau})^{\alpha+1}}, \qquad i = 1, \ldots, n, \tag{3.1}$$

for $d_i, \alpha, \tau, \lambda > 0$, where α and τ represent shape parameters, λ is a scale parameter (Hogg and Klugman, 1984) and n is the number of claim delays.

Here we consider an alternative Burr parameterisation (Dutang *et al.*, 2008), by defining a new scale parameter $s = \lambda^{1/\tau}$. This parameterisation was chosen mainly in order to improve the efficiency of the MCMC estimation algorithm. The density function is now given as

$$f(d_i; \alpha, \tau, s) = \frac{\alpha\tau(d_i/s)^{\tau}}{d_i[1 + (d_i/s)^{\tau}]^{\alpha+1}}, \qquad i = 1, \ldots, n,$$

for $d_i, \alpha, \tau, s > 0$. Under this setting, the distribution function is

$$F_D(d) = 1 - \left(\frac{s^\tau}{s^\tau + d^\tau}\right)^\alpha$$

with k^{th} moment given by

$$E(D^k) = s^k \Gamma\left(\alpha - \frac{k}{\tau}\right) \Gamma\left(1 + \frac{k}{\tau}\right) / \Gamma(\alpha), \tag{3.2}$$

for $\alpha\tau > k$, or ∞ otherwise.

In a GL-type setting, we regress covariates of interest on the mean of the distribution and use a logarithmic link since the mean duration should be strictly positive. This gives

$$\mathrm{E}(D_i) = \exp(\eta_i) \qquad i = 1, \ldots, n \tag{3.3}$$

where $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$, with $\boldsymbol{\eta}$ $n \times 1$, $\boldsymbol{X}$ the $n \times (p+1)$ design matrix and $\boldsymbol{\beta}$ denoting the $(p+1) \times 1$ coefficient vector. Here p is the number of covariates which are given later in Table 3.3.

To associate the mean of the delay distribution with the explanatory variables, we consider (3.2) and use the link function in (3.3) to obtain the Burr GL-type model

$$D_i \sim Burr(\alpha, \tau, s_i)$$

with

$$\begin{aligned} \mathrm{E}(D_i) &= \exp(\eta_i) \\ &= s_i \frac{\Gamma(\alpha - 1/\tau)\Gamma(1 + 1/\tau)}{\Gamma(\alpha)} \end{aligned} \tag{3.4}$$

which implies that

$$s_i = \frac{\Gamma(\alpha)}{\Gamma\left(\alpha - \frac{1}{\tau}\right)\Gamma\left(1 + \frac{1}{\tau}\right)} \exp(\eta_i).$$

Note that $s_i = \lambda_i^{1/\tau}$. To facilitate the fitting of this non-standard model, we consider expressing the Burr distribution as a mixed hierarchical model comprising a Weibull and a gamma component. This is particularly useful as the WinBUGS software (used

here for MCMC estimation) does not include the Burr distribution in its standard models. An alternative option, to explicitly define the Burr likelihood function in the WinBUGS code, proved to be very inefficient (Spiegelhalter *et al.*, 2003) and led to computational problems especially when we introduced missing values to the problem. Using the parameterisation given in (3.1), the mixed model is given as

$$\begin{aligned} D_i|\theta_i &\sim Weibull(\tau, \theta_i) \\ \theta_i &\sim Gamma(\alpha, \lambda_i) \end{aligned} \tag{3.5}$$

for $i = 1, \ldots, n$ and τ, θ_i, α, $\lambda_i > 0$. This can be shown by first noticing that the marginal density function of d_i under (3.5) is

$$f(d_i) = \int_0^\infty f(d_i|\theta_i)h(\theta_i)d\theta_i$$

with $f(.)$ and $h(.)$ being the probability density functions of d_i and θ_i. Here

$$f(d_i|\theta_i) = \tau\theta_i d_i^{\tau-1}\exp(-\theta_i d_i^\tau) \text{ and}$$

$$h(\theta_i) = \frac{\lambda_i^\alpha \theta_i^{\alpha-1}\exp(-\lambda_i\theta_i)}{\Gamma(\alpha)}.$$

This gives

$$f(d_i) = \frac{\tau d_i^{\tau-1}\lambda_i^\alpha}{\Gamma(\alpha)}\int_0^\infty \theta_i^\alpha \exp(-\theta_i(\lambda_i + d_i^\tau))d\theta_i$$

and by integrating over θ_i

$$f(d_i) = \frac{\tau d_i^{\tau-1}\lambda_i^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha+1)}{(\lambda_i + d_i^\tau)^{\alpha+1}}$$

which implies that the marginal density of d_i is

$$f(d_i) = \frac{\alpha\tau\lambda_i^\alpha d_i^{\tau-1}}{(\lambda_i + d_i^\tau)^{\alpha+1}}$$

i.e. $d_i|\alpha, \tau, \lambda_i \sim$ Burr$(\alpha, \tau, \lambda_i)$.

In the following sections we consider prior distributions for the model parameters under the Bayesian model. For the classical analysis, we maximised the following

log-likelihood function

$$l = \log(f(\boldsymbol{D}|\alpha,\tau,\beta)) = n\log(\alpha) + n\log(\tau) + \alpha\sum_i \log(\lambda_i) + (\tau-1)\sum_i \log(d_i) - (\alpha+1)\sum_i \log(\lambda_i + d_i^{\tau}). \tag{3.6}$$

The LN model

The probability model used for a lognormal GLM is

$$\begin{aligned} D_i &\sim LN(\mu_i, \sigma^2) \qquad i = 1,\ldots,n \\ \boldsymbol{\mu} &= \boldsymbol{\eta} = \boldsymbol{X\beta} \end{aligned} \tag{3.7}$$

where $\boldsymbol{\mu}$ being $n \times 1$ vector. The log-likelihood is

$$l = \log(f(\boldsymbol{D}|\sigma^2,\beta)) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \sum_i \log(d_i) - \frac{\sum_i (\log(d_i)-\mu_i)^2}{2\sigma^2}.$$

3.3.1 The null model

First consider the null model, i.e. the model with

$$\mathrm{E}(D_i) = \exp(\eta_i) = \exp(\eta) = \exp(\beta_0), \qquad i = 1,\ldots,15860.$$

This is equivalent to the GL-type model in (3.4) with

$$\mathrm{E}(D_i) = s\,G(\alpha,\tau)$$

where $G(\alpha,\tau) = \frac{\Gamma(\alpha-1/\tau)\Gamma(1+1/\tau)}{\Gamma(\alpha)}$. By setting $s = (G(\alpha,\tau))^{-1}\exp(\beta_0)$ we have

$$\beta_0 = \log(s) + \log(G(\alpha,\tau)).$$

In order for the mean $\mathrm{E}(D_i)$ to be defined, we need $\alpha\tau > 1$ and therefore we impose a restriction on the prior distribution of τ. Censoring is denoted here using the notation

$I(lower, upper)$. We therefore assign the following non-informative prior distributions

$$\alpha \sim Gamma(0.001, 0.001)$$

$$\tau \sim Gamma(0.001, 0.001)\ \mathrm{I}\left(\frac{1}{\alpha}, \infty\right)$$

$$s \sim Gamma(1, 0.01).$$

The likelihood of this model is

$$f(\boldsymbol{D}|\alpha, \tau, \boldsymbol{\beta}) = \alpha^n \tau^n \prod_i^n \left((G(\alpha,\tau))^{-1} \exp(\beta_0)\right)^{\alpha\tau} d_i^{\tau-1} \left[\left((G(\alpha,\tau))^{-1} \exp(\beta_0)\right)^{\tau} + d_i^{\tau}\right]^{-(\alpha+1)}. \tag{3.8}$$

Model parameters are estimated by using MCMC methodology in WinBUGS and by maximising the likelihood given in (3.8) in R software. To obtain the Bayesian estimates, 10000 iterations were performed after 4000 burn-in values. Chain traces and low MC errors showed that convergence was satisfied after burn-in.

Table 3.1 shows posterior estimates and ML estimates of the model parameters and also the mean, median, standard deviation and the log-likelihood value of the fitted Burr and lognormal distributions. Bayesian estimation and classical analysis give very close results. The mean delay under the Burr distribution is longer than with the lognormal model whereas the median delay of the fitted Burr distribution is shorter. This is because the Burr distribution is more skewed to the right due to its longer tail. For the Burr distribution, the condition of the existence of the second moment given in (3.2) is not satisfied, i.e. $\alpha\tau \not> 2$. So, the second moment of this distribution is not defined, giving infinite variance.

Also note that the estimate of τ under the Burr distribution (around 2.3 with 0.028 standard deviation) shows that the parameter is significantly greater than 1. Since $Burr(\alpha, 1, \lambda) \equiv Pareto(\alpha, \lambda)$, the estimate of τ suggests that a Pareto distribution would not be appropriate.

Table 3.1: Posterior and ML estimation under the null model.

	Burr					**LN**			
	Mean MCMC	SD MCMC	Mean MLE	SD MLE		Mean MCMC	SD MCMC	Mean MLE	SD MLE
α	0.7784	0.0191	0.7764	0.0193	μ	4.7516	0.0075	4.7516	0.0075
τ	2.2290	0.0282	2.2324	0.0275	σ	0.9434	0.0053	0.9434	0.0053
s	95.7123	1.6505	95.5588	1.6867					
Mean	193.26	2.9842	193.21	-		180.67	1.6202	180.66	-
Median	112.59	0.7548	112.59	-		115.77	0.8640	115.77	-
SD	∞	-	∞	-		216.47	3.3465	216.40	-
LogL	-96135.3		-96133.8			-96941.4		-96940.4	

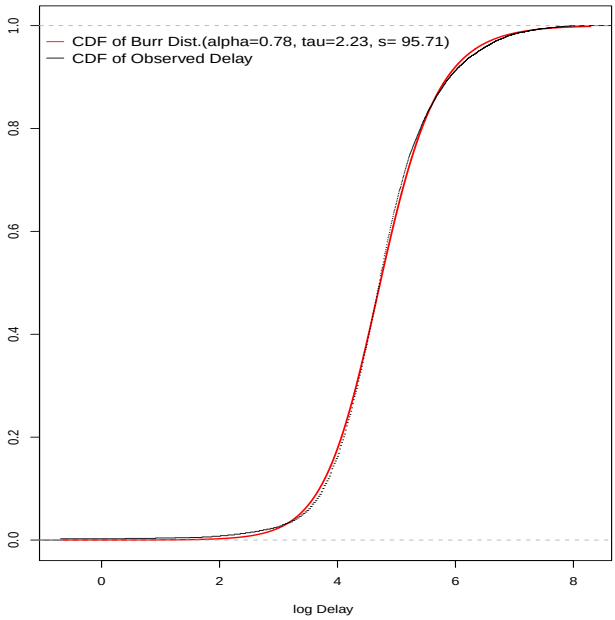
To compare the Burr and lognormal models, we use the deviance information criterion (DIC) under the Bayesian methodology (Spiegelhalter *et al.*, 2002) and Bayesian information criterion (BIC) under the classical analysis (Schwarz, 1978). To calculate the DIC, the effective number of parameters, p_D is used. This is given as the difference between the posterior mean of the deviance, $\bar{D} = -2\widehat{\log L}(\boldsymbol{\theta})$, and the deviance calculated at posterior means of the parameters, $\hat{D} = -2\log L(\hat{\boldsymbol{\theta}})$. Then, DIC $= \bar{D} + p_D$. BIC can be calculated as BIC $= -2\log L(\hat{\boldsymbol{\theta}}) + k\log(n)$ where k is the number of estimated parameters. The lower criteria values of the Burr distribution (see Table 3.2) imply a better fit of this model when covariates are not considered.

Table 3.2: Information criteria under the null model.

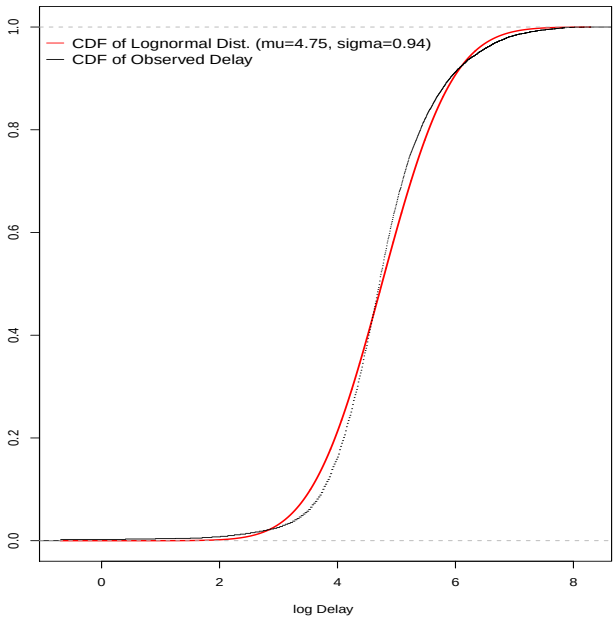
	Burr	**LN**
BIC (MLE)	192296.5	193900.1
$\bar{D} = -2\widehat{\log L}(\boldsymbol{\theta})$	192270.53	193882.86
$\hat{D} = -2\log L(\hat{\boldsymbol{\theta}})$	192267.51	193880.82
p_D	3.02	2.04
DIC (MCMC)	192273.55	193884.90

Figure 3.3 shows a comparison of the empirical cdf of observed duration and the fitted distribution using posterior estimates of the parameters under the Burr and lognormal distributions. When data are not subdivided, the cdf of the observed duration is very close to the cdf of the fitted Burr distribution. All comparisons of cdfs, log-likelihood values and information criteria suggest that the Burr distribution gives a considerably

better fit than the lognormal distribution.



(a) Burr Distribution



(b) Lognormal Distribution

Figure 3.3: CDF of the diagnosis – settlement interval.

3.3.2 Analysis with covariates

To model the delay, 10 covariates are used, namely age, sex, benefit type, smoker status, joint/single life, settlement year, benefit amount, policy duration, office and cause of claim. Details of these covariates are given in Table 3.3. The office covariate with 13 levels and cause of claim with 10 levels are taken as categorical variables. To compare office and cause effects with the corresponding mean level effect, a sum-to-zero constraint is used for these variables, whereas standardised values are used for other variables.

Table 3.3: Definitions of the covariates.

	Covariate	Number of Levels	Additional Information
x_1	Age	Numerical	age last birthday at date of diagnosis
x_2	Sex	2 (F & M)	F is the base category
x_3	Benefit type	2 (FA & SA)	FA is the base category
x_4	Smoker status	2 (N & S)	N is base category
x_5	Policy type	2 (Joint/Single life)	J is the base category
x_6	Settlement year	Numerical	record year
x_7	Benefit amount	Numerical	
x_8	Policy duration	Numerical	duration between dates of diagnosis and commencement of the policy
x_9	Office	13	
x_{10}	Cause of claim	10	1.CABG 2.Cancer 3.Death 4.Heart Attack 5.Kidney Failure 6.Major Organ Transplant 7.Multiple Sclerosis 8.Other 9.Stroke 10.Total and Permanent Disability

The Burr model

We regress the covariates on the mean of the Burr distribution and use the logarithmic link in (3.4) where

$$\eta_i = \beta_0 + \sum_{j=1}^{8} \beta_j z_{ij} + \beta_{9,Office_i} + \beta_{10,Cause_i} \tag{3.9}$$

for $i = 1, \ldots, 15860, \text{Office}_i = 1, \ldots, 13, \text{Cause}_i = 1, \ldots, 10$.

Here $z_{ij} = (x_{ij} - \bar{x}_j)/sd(x_j)$ are the standardised observations with x_{ij} being the original observations, β_0 and β_j $(j = 1, \ldots, 8)$ are the standardised intercept term and coefficients respectively and $\beta_{9,1}, \ldots, \beta_{9,13}$, $\beta_{10,1}, \ldots, \beta_{10,10}$ are the coefficients corresponding to the different levels of the office and cause factors. The coefficients on the original scale for the linear covariates, b_j, and the intercept term, b_0, can be calculated by $b_j = \beta_j/sd(x_j)$ and $b_0 = \beta_0 - \sum_{j=1}^{8} b_j \bar{x}_j$.

We assign the following non-informative prior distributions to the model parameters.

$$\begin{aligned}
&\alpha \sim Gamma(0.001, 0.001) \\
&\tau \sim Gamma(0.001, 0.001)\ \mathrm{I}(\frac{1}{\alpha}, \infty) \\
&\beta_j \sim N(0, 1000), j = 0, \ldots, 8 \\
&\beta_{9,k} \sim N(0, 1000), k = 2, \ldots, 13 \\
&\beta_{10,l} \sim N(0, 1000), l = 2, \ldots, 10 \\
&\text{with } \sum_{k=1}^{13} \beta_{9,k} = 0 \text{ and } \sum_{l=1}^{10} \beta_{10,l} = 0.
\end{aligned} \tag{3.10}$$

The likelihood of this model can be given as

$$f(\boldsymbol{D}|\alpha, \tau, \boldsymbol{\beta}) = \alpha^n \tau^n \prod_i^n \left((G(\alpha, \tau))^{-1} \exp(\eta_i)\right)^{\alpha\tau} d_i^{\tau-1} \left[\left((G(\alpha, \tau))^{-1} \exp(\eta_i)\right)^{\tau} + d_i^{\tau}\right]^{-(\alpha+1)}$$

where η_i is in (3.9) and the joint posterior density has the form

$$p(\alpha, \tau, \boldsymbol{\beta}|\boldsymbol{D}) \propto f(\boldsymbol{D}|\alpha, \tau, \boldsymbol{\beta})\pi(\alpha)\pi(\tau)\pi(\boldsymbol{\beta}) \tag{3.11}$$

where $\pi(\alpha)$, $\pi(\tau)$, $\pi(\boldsymbol{\beta})$ are the prior densities given in (3.10).

Posterior estimates and the ML estimates of the model parameters are presented in Table 3.4. The graphical representation of the posterior estimates with their 95% credible intervals can be seen in Figure 3.6. Bayesian estimates are obtained after

34000 iterations where the first 4000 iterations are considered as a burn-in process. MC errors for all estimated parameters were relatively small and are not presented here. Convergence of the algorithm was also checked by inspecting the chain trace of the parameters, which indicated that the Markov chains converged after 4000 iterations. ML estimates of the mean and standard deviation of the coefficients are very close to the Bayesian estimates. This is because we have a large amount of data and we use uninformative priors for the model parameters. Note that after including the covariates in the model, estimated means, standard deviations and 95% credible interval of τ suggest once again that τ is significantly different from one and thus a Pareto distribution does not fit the data.

Based on this model, the settlement year, policy duration, office, death and stroke have stronger effects on the delay. The data suggest that the delay between diagnosis and settlement in CII is shorter for stand alone policyholders (β_3) and smokers (β_4) whereas it is longer for younger ages (β_1), single life policyholders (β_5) and for more recent settlement years (β_6). The benefit amount (β_7) has a negative effect on the delay. The delay gets shorter as the time since the policy was effected gets longer (β_8). Office (β_9) also affects the time between diagnosis and settlement with different administrative procedures changing the diagnosis – settlement period among offices. The cause of claim (β_{10}) has a big impact on the delay. For CABG or cancer claims the delay is shorter than average while the shortest delay is associated to death claims. The definition of multiple sclerosis, given by the ABI, states that the symptoms of this disease should be persistent for a continuous period of at least 6 months (ABI, 2006). When there is this kind of waiting period, the delay until the date of settlement increases significantly. Other diseases such as stroke or major organ transplant also have a positive effect on the length of the delay. Note that the negative estimate of coefficient β_2 implies a shorter delay for males. However, the 95% credible interval suggests that there is high posterior probability that there is no effect of gender on the delay distribution.

Table 3.4: Coefficients of the Burr model without growth rate.

	MCMC					**MLE**	
Parameter	Mean	SD	2.5%	50%	97.5%	Mean	SD
β_0	5.2970	0.0289	5.2430	5.2970	5.3560	5.2938	0.0280
β_1	-0.0207	0.0073	-0.0347	-0.0207	-0.0065	-0.0205	0.0070
β_2	-0.0131	0.0066	-0.0256	-0.0132	0.0001	-0.0129	0.0067
β_3	-0.0285	0.0060	-0.0403	-0.0284	-0.0166	-0.0284	0.0061
β_4	-0.0190	0.0064	-0.0317	-0.0189	-0.0062	-0.0189	0.0064
β_5	0.0329	0.0063	0.0203	0.0328	0.0454	0.0333	0.0063
β_6	0.1149	0.0075	0.1000	0.1149	0.1293	0.1152	0.0073
β_7	-0.0373	0.0066	-0.0503	-0.0371	-0.0241	-0.0375	0.0066
β_8	-0.1164	0.0077	-0.1313	-0.1166	-0.1011	-0.1168	0.0078
$\beta_{9,1}$	0.2352	0.0245	0.1866	0.2354	0.2828	0.2365	0.0239
$\beta_{9,2}$	0.1287	0.0234	0.0826	0.1291	0.1741	0.1295	0.0218
$\beta_{9,3}$	-0.2016	0.0605	-0.3249	-0.1996	-0.0853	-0.1944	0.0611
$\beta_{9,4}$	0.1271	0.0521	0.0276	0.1284	0.2263	0.1313	0.0495
$\beta_{9,5}$	-0.1365	0.0369	-0.2085	-0.1369	-0.0642	-0.1402	0.0368
$\beta_{9,6}$	-0.5209	0.0818	-0.6747	-0.5255	-0.3627	-0.5338	0.0830
$\beta_{9,7}$	-0.2960	0.1282	-0.5191	-0.3150	-0.0177	-0.2997	0.1223
$\beta_{9,8}$	0.0733	0.0233	0.0238	0.0745	0.1159	0.0748	0.0217
$\beta_{9,9}$	-0.2040	0.0275	-0.2599	-0.2034	-0.1532	-0.2023	0.0267
$\beta_{9,10}$	0.2224	0.0315	0.1584	0.2229	0.2826	0.2245	0.0328
$\beta_{9,11}$	-0.0836	0.0212	-0.1261	-0.0833	-0.0434	-0.0825	0.0196
$\beta_{9,12}$	0.1927	0.0260	0.1424	0.1929	0.2451	0.1948	0.0259
$\beta_{9,13}$	0.4632	0.0772	0.3114	0.4629	0.6166	0.4613	0.0777
$\beta_{10,1}$	-0.1325	0.0418	-0.2144	-0.1327	-0.0513	-0.1318	0.0414
$\beta_{10,2}$	-0.0895	0.0194	-0.1267	-0.0898	-0.0506	-0.0889	0.0205
$\beta_{10,3}$	-0.4787	0.0275	-0.5324	-0.4791	-0.4247	-0.4784	0.0281
$\beta_{10,4}$	0.0256	0.0242	-0.0209	0.0254	0.0743	0.0259	0.0245
$\beta_{10,5}$	0.0762	0.0818	-0.0841	0.0775	0.2334	0.0832	0.0803
$\beta_{10,6}$	0.2541	0.1232	0.0038	0.2562	0.4950	0.2445	0.1218
$\beta_{10,7}$	0.1031	0.0346	0.0360	0.1028	0.1720	0.1032	0.0343
$\beta_{10,8}$	0.0156	0.0284	-0.0404	0.0162	0.0708	0.0163	0.0286
$\beta_{10,9}$	0.2412	0.0287	0.1854	0.2411	0.2973	0.2414	0.0290
$\beta_{10,10}$	-0.0150	0.0566	-0.1250	-0.0137	0.0940	-0.0155	0.0604
α	0.6200	0.0150	0.5917	0.6195	0.6513	0.6176	0.0168
τ	2.6290	0.0333	2.5580	2.6280	2.6920	2.6371	0.0383

Model fit under the Burr distribution

To explore the model fit, residuals can be used giving the discrepancy between actual data and model fitted values. For model adequacy, deviance residuals, r_D, are calculated as follows

$$r_{D_i} = \text{sign}(d_i - \hat{d}_i)\sqrt{2(l_{d_i} - l_{\hat{d}_i})}$$

where d and $\hat{d}$ are the actual and fitted delays, respectively with l being the log-likelihood value corresponding to them (McCullagh and Nelder, 1989).

Under the parameterisation given in (3.1), both α and τ are the same for each observation in the saturated likelihood but λ is allowed to vary with observations (see (3.6)). So the saturated log-likelihood is maximised when $dl/d\lambda_i = 0$. This gives

$$\frac{dl}{d\lambda_i} = \frac{\alpha}{\lambda_i} - \frac{\alpha + 1}{\lambda_i + d_i^\tau}. \tag{3.12}$$

Equating (3.12) to 0, we get

$$\lambda_i = \alpha d_i^\tau \tag{3.13}$$

The deviance can be obtained by substituting (3.13) into the saturated log-likelihood

$$l_{d_i} = \log(\alpha) + \log(\tau) + \alpha \log(\alpha d_i^\tau) + (\tau - 1)\log(d_i) - (\alpha + 1)\log(\alpha d_i^\tau + d_i^\tau)$$

$$l_{\hat{d}_i} = \log(\alpha) + \log(\tau) + \alpha \log(\alpha \hat{d}_i^\tau) + (\tau - 1)\log(d_i) - (\alpha + 1)\log(\alpha \hat{d}_i^\tau + d_i^\tau)$$

and subtracting one from the other gives

$$l_{d_i} - l_{\hat{d}_i} = \alpha\tau \log(d_i/\hat{d}_i) - (\alpha + 1)\log\left(\frac{\alpha d_i^\tau + d_i^\tau}{\alpha \hat{d}_i^\tau + d_i^\tau}\right). \tag{3.14}$$

A plot of deviance residuals against the fitted delays is given in Figure 3.4. For convenience of plotting, logarithms of the fitted delays are shown in the figure. The graph shows no significant patterns in deviance residuals. The straight lines at the bottom of the figure correspond to the claims with 0.5, 1.5 and 2.5 days delay. This is both because of the log-linear structure of the model and the discrete nature of the response variable.

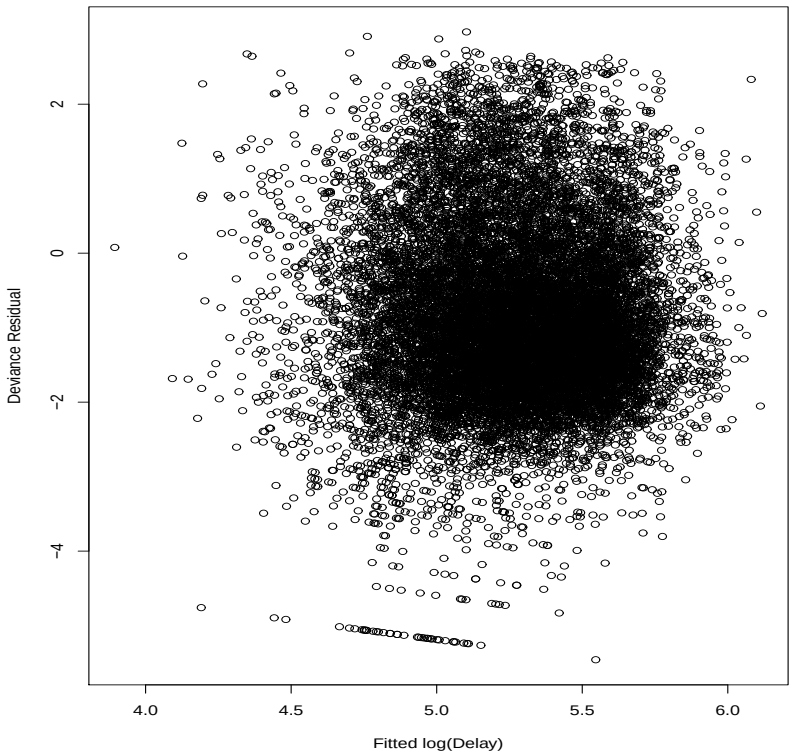


Figure 3.4: Deviance residuals of the Burr model without growth rate.

The LN model

The probability model used for a lognormal GLM is given as

$$\begin{aligned} D_i &\sim LN(\mu_i, \sigma^2) \\ \mu_i = \eta_i &= \beta_0 + \sum_{j=1}^{8} \beta_j z_{ij} + \beta_{9,Office_i} + \beta_{10,Cause_i}, \qquad i = 1, ..., 15860 \end{aligned} \tag{3.15}$$

with prior distributions for the β coefficients given in (3.10) and

$$\sigma^2 \sim IGa(0.001, 0.001).$$

Here IGa refers to 'Inverse Gamma' distribution. Therefore the likelihood can be

expressed as

$$f(\boldsymbol{D}|\sigma^2,\boldsymbol{\beta}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_i^n \frac{1}{d_i} \exp\left(-\frac{\sum(\log(d_i)-\eta_i)^2}{2\sigma^2}\right),$$

and the joint posterior density is

$$p(\sigma^2,\boldsymbol{\beta}|\boldsymbol{D}) \propto f(\boldsymbol{D}|\sigma^2,\boldsymbol{\beta})\pi(\sigma^2)\pi(\boldsymbol{\beta}).$$

where $\pi(\sigma^2)$ is the gamma prior density for σ^2. Posterior and ML estimates of the model parameters are given in Table 3.5 for the LN model. Posterior estimates and their 95% credible intervals are graphically represented in Figure 3.6. The two methods give very close estimates for the mean and standard deviation of all the coefficients. Here we note that MCMC convergence under this model is satisfied faster than for the Burr model.

Under the LN model, non-smokers (β_4), younger ages (β_1), single life policyholders (β_5) face longer diagnosis – settlement delays. This duration is shorter for the policyholders who have stand alone policies (β_3) and for earlier settlement years (β_6). Increasing benefit amount (β_7) and longer policy duration (β_8) leads to shorter delays until settlement. Office 4 ($\beta_{9,4}$) has the longest delay whereas Office 6 ($\beta_{9,6}$) has the shortest. Major organ transplant ($\beta_{10,6}$) has the strongest positive effect on the delay among causes. Since we have fewer data for this cause, it has a wide credible interval. Stroke comes next ($\beta_{10,9}$) with a shorter credible interval. The only covariate which has high posterior probability that there is no effect on the delay distribution is gender under this model.

Table 3.5: Coefficients of the LN model without growth rate.

			MCMC			**MLE**	
Parameter	Mean	SD	2.5%	50%	97.5%	Estimate	SD
β_0	4.7660	0.0273	4.7120	4.7660	4.8180	4.7628	0.0273
β_1	-0.0294	0.0081	-0.0454	-0.0294	-0.0133	-0.0294	0.0081
β_2	-0.0097	0.0077	-0.0249	-0.0097	0.0055	-0.0096	0.0077
β_3	-0.0389	0.0073	-0.0534	-0.0389	-0.0248	-0.0390	0.0073
β_4	-0.0256	0.0074	-0.0400	-0.0257	-0.0112	-0.0255	0.0074
β_5	0.0315	0.0073	0.0171	0.0315	0.0459	0.0315	0.0074
β_6	0.1272	0.0084	0.1106	0.1271	0.1434	0.1273	0.0084
β_7	-0.0509	0.0073	-0.0652	-0.0509	-0.0365	-0.0509	0.0073
β_8	-0.1755	0.0088	-0.1926	-0.1755	-0.1583	-0.1756	0.0087
$\beta_{9,1}$	0.1688	0.0291	0.1124	0.1686	0.2270	0.1701	0.0293
$\beta_{9,2}$	0.0763	0.0262	0.0250	0.0759	0.1281	0.0779	0.0266
$\beta_{9,3}$	-0.1772	0.0690	-0.3099	-0.1774	-0.0406	-0.1791	0.0687
$\beta_{9,4}$	0.3469	0.0509	0.2472	0.3464	0.4477	0.3492	0.0517
$\beta_{9,5}$	-0.1668	0.0439	-0.2538	-0.1666	-0.0820	-0.1655	0.0436
$\beta_{9,6}$	-0.7012	0.0889	-0.8694	-0.7038	-0.5217	-0.6956	0.0866
$\beta_{9,7}$	-0.4013	0.1661	-0.7349	-0.4026	-0.0732	-0.4180	0.1629
$\beta_{9,8}$	0.0397	0.0263	-0.0120	0.0397	0.0916	0.0409	0.0265
$\beta_{9,9}$	-0.0709	0.0308	-0.1308	-0.0711	-0.0101	-0.0693	0.0305
$\beta_{9,10}$	0.2475	0.0399	0.1698	0.2473	0.3253	0.2494	0.0398
$\beta_{9,11}$	0.0005	0.0235	-0.0448	0.0000	0.0477	0.0022	0.0235
$\beta_{9,12}$	0.3048	0.0297	0.2468	0.3047	0.3633	0.3067	0.0296
$\beta_{9,13}$	0.3330	0.0978	0.1288	0.3361	0.5223	0.3311	0.1027
$\beta_{10,1}$	-0.1365	0.0477	-0.2306	-0.1361	-0.0438	-0.1355	0.0475
$\beta_{10,2}$	-0.0939	0.0218	-0.1369	-0.0936	-0.0520	-0.0922	0.0223
$\beta_{10,3}$	-0.5986	0.0296	-0.6577	-0.5984	-0.5411	-0.5975	0.0299
$\beta_{10,4}$	0.0031	0.0270	-0.0509	0.0035	0.0551	0.0043	0.0274
$\beta_{10,5}$	0.1283	0.0865	-0.0425	0.1295	0.2965	0.1272	0.0851
$\beta_{10,6}$	0.2631	0.1350	0.0040	0.2616	0.5315	0.2536	0.1407
$\beta_{10,7}$	0.0497	0.0369	-0.0227	0.0498	0.1220	0.0512	0.0368
$\beta_{10,8}$	-0.0164	0.0317	-0.0783	-0.0163	0.0449	-0.0145	0.0321
$\beta_{10,9}$	0.2205	0.0327	0.1557	0.2208	0.2841	0.2217	0.0327
$\beta_{10,10}$	0.1807	0.0436	0.0969	0.1802	0.2674	0.1818	0.0440
σ^2	0.7981	0.0090	0.7807	0.7980	0.8159	0.7979	

The plot of deviance residuals against the logarithms of the fitted delays is given in Figure 3.5. The graph shows no significant patterns in deviance residuals. The straight lines at the bottom appear under the lognormal model corresponding to 0.5, 1.5 and 2.5 days of delay as well. This is, again, due to the log-linear structure of the model and the discrete nature of the response variable. Note that 3.8% of the residuals lie outside [-2,2].

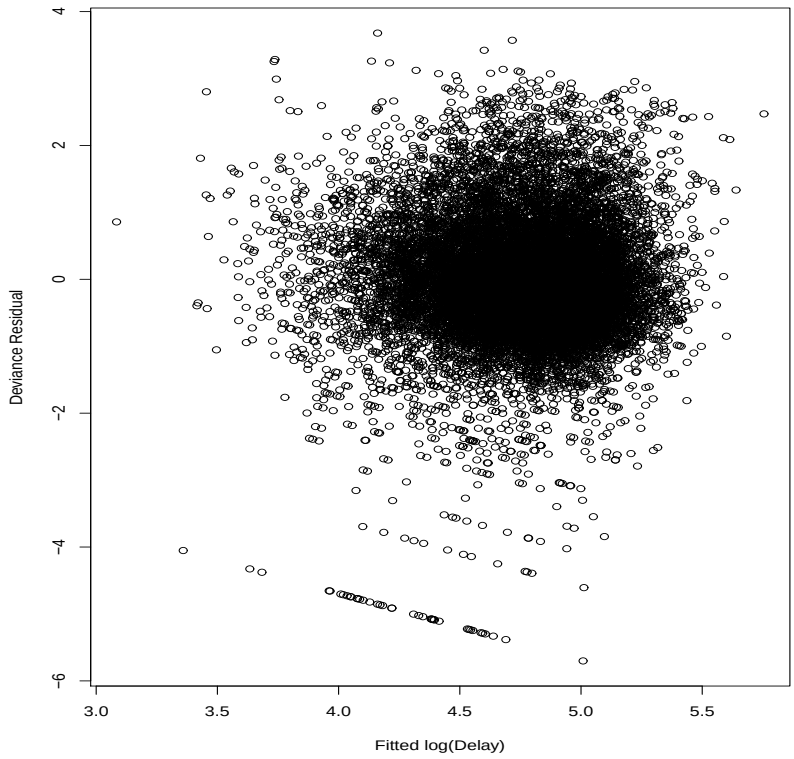


Figure 3.5: Deviance residuals of the LN model without growth rate.

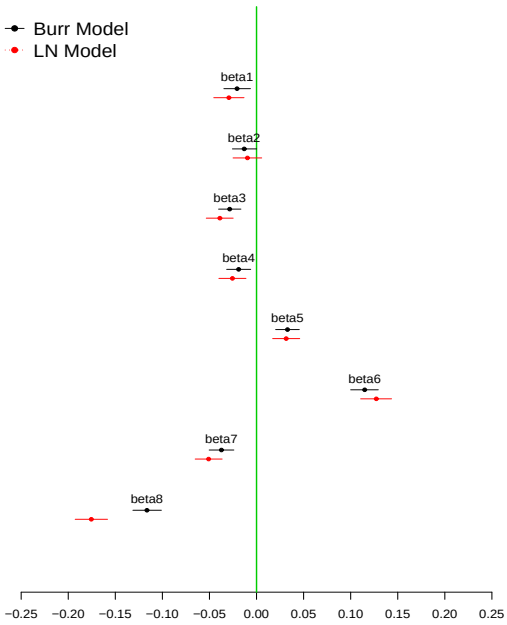
Comparison of Burr and LN models

To compare the Burr and lognormal models, the corresponding estimates of DIC and BIC are given in Table 3.6. The lower values of information criteria of the Burr model suggests that the fit is better under this distribution.

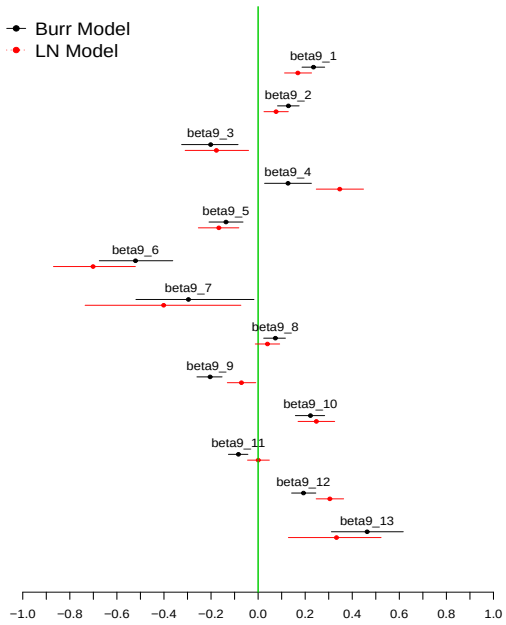
Table 3.6: Values of information criteria of the models without growth rate.

	$\bar{D} = -2\widehat{\log L}(\boldsymbol{\theta})$	$\hat{D} = -2\log L(\hat{\boldsymbol{\theta}})$	p_D	**DIC**	**BIC**
Burr	190348.2	190316.7	31.5	190379.7	190626.0
LN	192150.8	192120.3	30.5	192181.3	192420.1

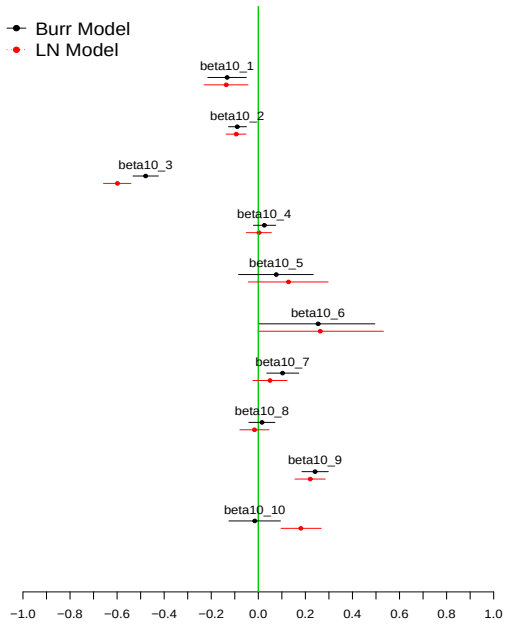
Figure 3.6 shows a comparison between the posterior estimates and 95% credible intervals of the coefficients under both models. All the coefficients have the same direction of effect on the delay when they are significant under the two models. However magnitudes of the effects are changing significantly for some coefficients, such as, policy duration (β_8), death ($\beta_{10,3}$), TPD ($\beta_{10,10}$) or Office 4, Office 9 and Office 12 ($\beta_{9,4}$, $\beta_{9,9}$ and $\beta_{9,12}$). The first two have stronger negative effect and the latter ones (except Office 9) have stronger positive effect under the LN model compared to the Burr model. The negative effect of Office 9 is stronger under the Burr distribution. These covariates/levels might be the most affected covariates by the shape (e.g. long tail) of the distributions.



(a) $\beta_1 - \beta_8$



(b) β_9



(c) β_{10}

Figure 3.6: Comparison of posterior estimates of the coefficients under Burr (black line) and LN (red line) model. Bars show 95% credible intervals and bullets show posterior means.

3.4 Prediction under the two models

We now consider 17 scenarios for different risk types, as shown in Table 3.7. Note that benefit amounts in the table are in thousands and the currency is GBP. The first scenario is formed using the average values of the numerical covariates and the levels of the factors where we have most of the data. The other scenarios show how the delay changes when claim-related factors are altered. The changing variables are shown in bold font in the table. Selected estimates of the predicted posterior delay of the scenarios are given in Table 3.8 under the Burr and LN model. Also, 95% credible intervals of the predicted posterior means of the scenarios are shown in Figure 3.7 for both models.

Posterior estimates of the means show that the mean delays are longer under the Burr distribution for almost all of the cases. Considering the longer tail of the distribution (see the upper bounds of the credible intervals in Table 3.8), this result is expected. Nevertheless these differences are not significant for some scenarios as the credible intervals of the means of the scenarios are overlapping substantially under the two models.

The mean delay of the typical scenario (Scenario 1) is 149 days under the Burr distribution and 140 days under the LN model (these are shown with vertical lines in Figure 3.7). For settlement year 1999 (2005) (Scenarios 8 and 9) these delays are changing to 123 days (181 days) under the Burr model and 113 days (173 days) under the LN model.

In the typical case, the policy duration is taken as 4 years. Increasing this period to 10 years (Scenario 13) decreases the mean delay to 114 days under the Burr model and 93 days under the LN model. The estimated delay is 101 days (95 days under the LN model) for the death claims (Scenario 16). These two scenarios (Scenarios 13 and 16) are also the only scenarios where the 95% credible intervals of the means under the two models do not overlap at all. For Scenario 17, holding the other covariates constant, for TPD claims the delay between dates of diagnosis and settlement is longer under the LN model (185 days) than the Burr model (161 days). However we note that the coefficient of TPD ($\beta_{10,10}$) is not significant under the Burr model. Scenarios 13, 16 and 17 are related to policy duration, death and TPD. In the previous section,

it is mentioned that these variables might be more sensitive to the structure of the distribution.

The credible intervals of the Scenarios 15 and 17 are wider than the others. Among these scenarios, Scenario 14 (for Office 6) gives the shortest mean delay with 95 days (70 days for the LN model).

Table 3.7: Scenarios for prediction of the CDD.

Scenario (Scn)	1	2	3	4	5	6	7	8	9
Sex	M	**F**	M	M	M	M	M	M	M
Benefit Type	FA	FA	**SA**	FA	FA	FA	FA	FA	FA
Smoker Status	NS	NS	NS	**S**	NS	NS	NS	NS	NS
Policy Type	J	J	J	J	**S**	J	J	J	J
Age	45	45	45	45	45	**20**	**65**	45	45
Settlement Year	2002	2002	2002	2002	2002	2002	2002	**1999**	**2005**
Benefit Amount	50	50	50	50	50	50	50	50	50
Policy Duration	1460	1460	1460	1460	1460	1460	1460	1460	1460
Office Code	11	11	11	11	11	11	11	11	11
Cause of Claim	Cancer	Cancer	Cancer	Cancer	Cancer	Cancer	Cancer	Cancer	Cancer

Scenario	10	11	12	13	14	15	16	17
Sex	M	M	M	M	M	M	M	M
Benefit Type	FA	FA	FA	FA	FA	FA	FA	FA
Smoker Status	NS	NS	NS	NS	NS	NS	NS	NS
Policy Type	J	J	J	J	J	J	J	J
Age	45	45	45	45	45	45	45	45
Settlement Year	2002	2002	2002	2002	2002	2002	2002	2002
Benefit Amount	**10**	**250**	50	50	50	50	50	50
Policy Duration	1460	1460	**365**	**3650**	1460	1460	1460	1460
Office Code	11	11	11	11	**6**	**10**	11	11
Cause of Claim	Cancer	Cancer	Cancer	Cancer	Cancer	Cancer	**Death**	**TPD**

Table 3.8: Posterior estimates of mean delays for the scenarios in Table 3.7 for the Burr and LN model (days).

	Burr Model					**LN Model**				
	Mean	SD	2.5%	50%	97.5%	Mean	SD	2.5%	50%	97.5%
Mean.Scn1	148.9	3.6	142.0	148.8	156.2	140.1	2.8	134.6	140.0	145.7
Mean.Scn2	152.8	3.6	145.8	152.7	160.2	142.8	2.7	137.5	142.8	148.3
Mean.Scn3	136.5	4.2	128.2	136.5	144.8	124.2	3.6	117.2	124.1	131.4
Mean.Scn4	142.5	3.8	135.4	142.5	150.2	132.1	3.1	126.2	132.1	138.3
Mean.Scn5	159.1	3.9	151.9	158.9	167.4	149.2	3.1	143.2	149.2	155.5
Mean.Scn6	157.3	5.1	147.5	157.2	167.6	151.4	4.7	142.6	151.3	160.8
Mean.Scn7	142.5	3.8	135.1	142.4	150.4	131.7	3.3	125.3	131.6	138.4
Mean.Scn8	122.5	3.5	116.0	122.4	129.3	113.1	3.0	107.4	113.1	119.2
Mean.Scn9	181.0	4.6	172.1	180.9	190.6	173.4	3.9	165.8	173.4	181.2
Mean.Scn10	152.9	3.7	145.9	152.9	160.6	145.2	3.0	139.3	145.1	151.2
Mean.Scn11	130.2	4.5	121.6	130.1	139.3	117.2	3.8	109.9	117.1	124.9
Mean.Scn12	170.4	4.0	162.7	170.4	178.8	171.6	3.5	164.8	171.6	178.6
Mean.Scn13	113.6	3.9	106.5	113.4	121.5	93.3	3.0	87.7	93.3	99.3
Mean.Scn14	94.6	9.1	75.5	94.6	111.7	69.8	6.8	58.0	69.2	84.3
Mean.Scn15	201.4	8.0	186.3	201.2	217.9	179.4	7.4	165.5	179.2	194.5
Mean.Scn16	101.1	3.2	95.0	101.0	107.6	84.6	2.5	79.7	84.6	89.5
Mean.Scn17	161.4	10.1	141.0	161.4	180.3	184.5	8.5	168.6	184.2	202.0

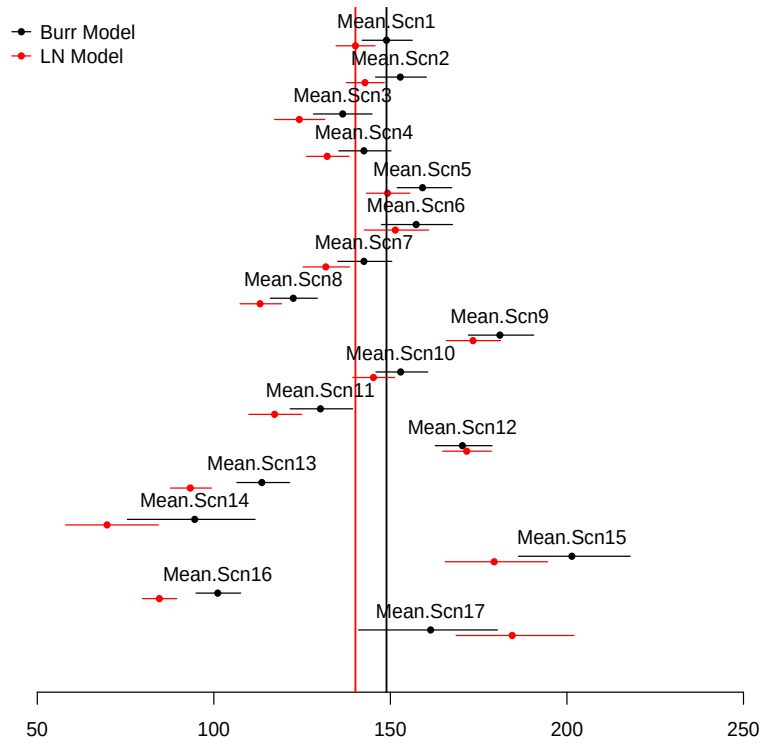


Figure 3.7: Comparison of posterior estimates of the mean delay under different scenarios using the Burr (black line) and LN model (red line). Bars show 95% credible intervals and bullets show posterior means. Vertical lines show the posterior means of the first scenarios under the two models.

Chapter 4

Modelling CDD II: Without considering the missing values and taking growth rates into account

4.1 Introduction

In this chapter, we relax the assumption of no growth rate within each office. Changes in the business are taken into consideration for each office. The CMI's data for CII is growing mainly because of two reasons; new offices and new business. Since the office codes are given in the data set, growth arising from new offices is handled in the model. On the other hand, growth within each office between successive years should be taken into account as this is the effect of new business. Table 4.1 shows growth rates (based on changes in average inforce figures) and growth factors for each office in each contribution year. As stated before, the data relate to claims where the settlement years are between 1999 and 2005. This is also the maximum period of the given inforce data for an office. If an office partly contributed within this period, e.g. from 1999 to 2002, the inforce data for that office is limited to that period. So, the average number of inforce is used to calculate the growth rate for each office between successive years. However, some of the claims' diagnosis dates lie before that office's first contribution year. For those claims, growth rates should be estimated from the known observations. Here, modelling the growth rates is not reasonable

because there is lack of observed growth rates for each office and they are generally very volatile. Rather, we preferred to use the earliest growth rate for the ones we should estimate. Based on this estimation, the average inforce and the growth factors can be calculated. In Table 4.1, the estimated figures are in bold font. Results, considering the growth rates, are presented in Section 4.2. In Section 4.3, prediction results are given for the same hypothetical scenarios introduced in Chapter 3. Model coefficients and prediction results are also compared when the growth rate is ignored or taken into account in the models.

Table 4.1: Growth rates from the inforce data for offices between successive years.

Office	Year	Growth Rate (GR)	Growth Factor (GF)	Office	Year	Growth Rate (GR)	Growth Factor (GF)
1	2001		**1.5390**	8	2000	0.9147	4.0000
1	2002	**0.1305**	**1.3613**	8	2001	0.6676	2.3987
1	2003	**0.1305**	1.2041	8	2002	0.7589	1.3638
1	2004	0.1305	1.0651	8	2003	0.2443	1.0960
1	2005	0.0651	1	8	2004	0.1460	0.9564
				8	2005	-0.0436	1
2	1997		**15.4865**				
2	1998	**0.4357**	**10.7870**	9	1993		**1.2275**
2	1999	**0.4357**	7.5136	9	1994	**0.0337**	**1.1875**
2	2000	0.4357	5.2336	9	1995	**0.0337**	**1.1488**
2	2001	0.8013	2.9055	9	1996	**0.0337**	**1.1113**
2	2002	0.7720	1.6396	9	1997	**0.0337**	**1.0751**
2	2003	0.3811	1.1872	9	1998	**0.0337**	**1.0400**
2	2004	0.1386	1.0426	9	1999	**0.0337**	**1.0061**
2	2005	0.0426	1	9	2000	**0.0337**	0.9733
				9	2001	0.0337	0.9416
3	1997		**14.3470**	9	2002	-0.0119	0.9530
3	1998	**0.8440**	**7.7804**	9	2003	-0.0276	0.9801
3	1999	**0.8440**	4.2193	9	2004	-0.0205	1.0006
3	2000	0.8440	2.2881	9	2005	0.0006	1
3	2001	0.5942	1.4353				
3	2002	0.4353	1	10	1997		**2.4990**
				10	1998	**0.1932**	**2.0944**
4	1994		**0.3618**	10	1999	**0.1932**	1.7554
4	1995	**-0.0499**	**0.3808**	10	2000	0.1932	1.4712
4	1996	**-0.0499**	**0.4008**	10	2001	0.1082	1.3276
4	1997	**-0.0499**	**0.4219**	10	2002	0.0964	1.2108
4	1998	**-0.0499**	**0.4441**	10	2003	0.0807	1.1204
4	1999	**-0.0499**	0.4674	10	2004	0.1235	0.9972
4	2000	-0.0499	0.4920	10	2005	-0.0028	1
4	2001	-0.1039	0.5490				
4	2002	-0.1438	0.6413	11	1995		**3.8674**
4	2003	-0.1503	0.7547	11	1996	**0.2193**	**3.1717**
4	2004	-0.1405	0.8780	11	1997	**0.2193**	**2.6011**
4	2005	-0.1220	1	11	1998	**0.2193**	**2.1332**
				11	1999	**0.2193**	1.7495
5	2000		6.0575	11	2000	0.2193	1.4348
5	2001	1.8617	2.1167	11	2001	0.1655	1.2311
5	2002	0.5098	1.4020	11	2002	0.1390	1.0808
5	2003	0.1988	1.1695	11	2003	0.0810	0.9998
5	2004	0.0970	1.0661	11	2004	0.0214	0.9788
5	2005	0.0661	1	11	2005	-0.0212	1
6	1998		**15.1923**	12	1993		**0.5080**
6	1999	**0.2089**	**12.5667**	12	1994	**0.0001**	**0.5079**
6	2000	**0.2089**	10.3949	12	1995	**0.0001**	**0.5078**
6	2001	0.2089	8.5985	12	1996	**0.0001**	**0.5077**
6	2002	0.7340	4.9586	12	1997	**0.0001**	**0.5077**
6	2003	0.7422	2.8462	12	1998	**0.0001**	**0.5076**
6	2004	0.7862	1.5934	12	1999	**0.0001**	**0.5075**
6	2005	0.5934	1	12	2000	**0.0001**	0.5074
				12	2001	0.0001	0.5074
7	2003		**6.4257**	12	2002	-0.1129	0.5719
7	2004	**1.5349**	2.5349	12	2003	-0.1875	0.7039
7	2005	1.5349	1	12	2004	-0.1730	0.8512
				12	2005	-0.1488	1
8	1994		**197.0800**				
8	1995	**0.9147**	**102.9308**	13	1999		**2.3283**
8	1996	**0.9147**	**53.7586**	13	2000	**0.3047**	1.7846
8	1997	**0.9147**	**28.0770**	13	2001	0.3047	1.3679
8	1998	**0.9147**	**14.6640**	13	2002	0.3679	1
8	1999	**0.9147**	7.6587				

4.2 Modelling when the business growth is taken into account

The Burr model

As we only have the settled claims in the data set, recent claims have shorter delays whereas earlier claims show the real delay structure with longer delays. However when the growth rate is positive, the earlier observations are under-represented as they are fewer in number. The average inforce numbers for the offices are changing across years. This means that the number of claims would have been different, if there had been the current amount of CI business in earlier years. Therefore we assign claims office-specific weights according to their year of diagnosis. For example for Office 10 claims are weighted by 0.9972 if their year of diagnosis is 2004. This is because of the negative growth (-0.28%) between 2004 and 2005. Likewise, the claims are given a weight equal to 2.4990 weights if they are diagnosed in 1997 for this office. All claims diagnosed in the latest year of contribution of an office have unit weights. For some offices, growth rates are very large for earlier years. However these are relatively small offices and the corresponding very large growth rates do not have a significant effect on the modelling.

It is known that for estimation weights can be used in the model through the variance of the estimates (e.g. as in common least squares estimation or weighted least squares), i.e. weights can be inversely proportional to the variance (Greene, 1990). The variance of the Burr distribution can be written in the following form

$$V(D) = s^2 \frac{\Gamma(\alpha)\Gamma(\alpha - 2/\tau)\Gamma(1 + 2/\tau) - (\Gamma(\alpha - 1/\tau))^2(\Gamma(1 + 1/\tau))^2}{(\Gamma(\alpha))^2}. \tag{4.1}$$

The growth factors (GF) in Table 4.1 show weights assigned to each office between successive years. To introduce weights in the variance

$$s_w = s/\sqrt{GF}$$

is used in the models.

So, the probability model for a Burr GL-type model without missing values can be given in the following way

$$D_i \sim Burr(\alpha, \tau, \lambda_{w_i}), \qquad i = 1, \ldots, 15860 \tag{4.2}$$

where $\lambda_{w_i} = (s_{w_i})^\tau$ with $s_{w_i} = s_i/\sqrt{GF_i}$ and $s_i = (G(\alpha,\tau))^{-1}\exp(\eta_i)$, $G(\alpha,\tau)$ is as defined in Section 3.3.1. Here we define GF_i as the GF for observation i. We also assign prior distributions

$$\begin{aligned}
&\alpha \sim Gamma(0.01, 0.01)\\
&\tau \sim Gamma(0.01, 0.01)\ \mathrm{I}\left(\frac{1}{\alpha}, \infty\right)\\
&\beta_j \sim N(0, 1000), j = 1, \ldots, 8\\
&\beta_{9,k} \sim N(0, 100), k = 2, \ldots, 13\\
&\beta_{10,l} \sim N(0, 100), l = 2, \ldots, 10\\
&\text{with } \sum_{k=1}^{13} \beta_{9,k} = 0 \text{ and } \sum_{l=1}^{10} \beta_{10,l} = 0.
\end{aligned} \tag{4.3}$$

The log-likelihood function of this model can be written in the following form

$$\begin{aligned}
l = \log(f(\boldsymbol{D}|\alpha,\tau,\beta)) =& n\log(\alpha) + n\log(\tau) + \alpha\sum_i \log(\lambda_{w_i}) + (\tau-1)\sum_i \log(d_i) -\\
&(\alpha+1)\sum_i \log(\lambda_{w_i} + d_i^\tau).
\end{aligned} \tag{4.4}$$

The joint density function has exactly the same form as in (3.11) where $\pi(\alpha)$, $\pi(\tau)$, $\pi(\boldsymbol{\beta})$ are the prior densities given in (4.3). By maximising (4.4), estimates of the standardised coefficients and their standard deviations can be obtained. These are given in Table 4.2 together with the posterior estimates. Posterior estimates are obtained by using MCMC. A total of 30000 iterations are performed after a 4000 iteration burn-in process. Here, we see again that the ML estimates of the parameters

are very close to those estimated using Bayesian methodology. The posterior estimates and their 95% credible intervals are graphically represented in Figure 4.4.

Table 4.2: Coefficients of the Burr model with growth rate.

	MCMC					**MLE**	
Parameter	Mean	SD	2.5%	50%	97.5%	Mean	SD
β_0	5.4400	0.0293	5.3840	5.4400	5.4960	5.4351	0.0304
β_1	-0.0200	0.0074	-0.0343	-0.0200	-0.0058	-0.0199	0.0072
β_2	-0.0132	0.0069	-0.0267	-0.0131	0.0000	-0.0127	0.0068
β_3	-0.0221	0.0063	-0.0344	-0.0220	-0.0100	-0.0221	0.0063
β_4	-0.0177	0.0065	-0.0307	-0.0178	-0.0045	-0.0175	0.0065
β_5	0.0374	0.0066	0.0244	0.0374	0.0503	0.0374	0.0065
β_6	0.0167	0.0077	0.0015	0.0169	0.0315	0.0169	0.0077
β_7	-0.0358	0.0066	-0.0488	-0.0358	-0.0231	-0.0354	0.0068
β_8	-0.0996	0.0078	-0.1152	-0.0995	-0.0845	-0.0994	0.0079
$\beta_{9,1}$	0.3021	0.0240	0.2555	0.3021	0.3494	0.3043	0.0246
$\beta_{9,2}$	0.2489	0.0224	0.2054	0.2486	0.2926	0.2510	0.0229
$\beta_{9,3}$	-0.1870	0.0627	-0.3081	-0.1865	-0.0641	-0.1794	0.0640
$\beta_{9,4}$	-0.2605	0.0518	-0.3604	-0.2603	-0.1581	-0.2616	0.0514
$\beta_{9,5}$	-0.0577	0.0393	-0.1347	-0.0576	0.0213	-0.0551	0.0380
$\beta_{9,6}$	-0.1582	0.0917	-0.3400	-0.1604	0.0168	-0.1544	0.0925
$\beta_{9,7}$	-0.0886	0.1297	-0.3522	-0.0843	0.1573	-0.1144	0.1258
$\beta_{9,8}$	0.1432	0.0216	0.1010	0.1429	0.1869	0.1455	0.0225
$\beta_{9,9}$	-0.3057	0.0275	-0.3593	-0.3058	-0.2514	-0.3042	0.0275
$\beta_{9,10}$	0.2305	0.0324	0.1672	0.2307	0.2944	0.2346	0.0337
$\beta_{9,11}$	-0.1064	0.0198	-0.1435	-0.1066	-0.0674	-0.1041	0.0203
$\beta_{9,12}$	-0.1691	0.0261	-0.2200	-0.1684	-0.1192	-0.1675	0.0268
$\beta_{9,13}$	0.4085	0.0764	0.2504	0.4109	0.5489	0.4054	0.0803
$\beta_{10,1}$	-0.1353	0.0423	-0.2182	-0.1350	-0.0531	-0.1340	0.0422
$\beta_{10,2}$	-0.0828	0.0199	-0.1214	-0.0825	-0.0446	-0.0806	0.0208
$\beta_{10,3}$	-0.4685	0.0275	-0.5249	-0.4685	-0.4152	-0.4682	0.0287
$\beta_{10,4}$	0.0248	0.0248	-0.0222	0.0247	0.0727	0.0265	0.0250
$\beta_{10,5}$	0.1100	0.0822	-0.0506	0.1112	0.2710	0.1067	0.0820
$\beta_{10,6}$	0.2497	0.1215	0.0166	0.2452	0.4911	0.2451	0.1228
$\beta_{10,7}$	0.1024	0.0345	0.0360	0.1017	0.1689	0.1049	0.0352
$\beta_{10,8}$	0.0210	0.0289	-0.0355	0.0207	0.0790	0.0227	0.0292
$\beta_{10,9}$	0.2479	0.0288	0.1898	0.2480	0.3047	0.2489	0.0296
$\beta_{10,10}$	-0.0692	0.0631	-0.1924	-0.0691	0.0526	-0.0720	0.0607
α	0.5879	0.0144	0.5596	0.5883	0.6142	0.5847	0.0158
τ	2.6170	0.0355	2.5570	2.6170	2.6970	2.6287	0.0387

The estimates of most of the coefficients are very similar when compared with the estimates under the assumption of no growth rate within the offices (see Figure 4.2 for the comparison between including and excluding growth rate in the Burr model). One of the most affected coefficients is the office (β_9), the other one is the settlement year (β_6). Positive effects of Office 4 ($\beta_{9,4}$) and Office 12 ($\beta_{9,12}$) on the delay distribution change to negative after we considered the business growth in the model. These two

offices are the only ones which are getting smaller during their contribution period. The strong positive effect of the settlement year on the delay distribution is significantly reduced. Since we allow for the business growth within each office between successive years, these results are expected.

A plot of deviance residuals against the logarithms of the fitted delays is given in Figure 4.1 to assess the model fit. From Table 4.1, it is seen that most of the offices have a positive growth rate during the period. Therefore, the claims diagnosed in earlier years have larger weights than the claims diagnosed in later years which means that they have smaller variances. This effect can be seen in Figure 4.1: the deviance residuals are getting slightly narrower for longer delays.

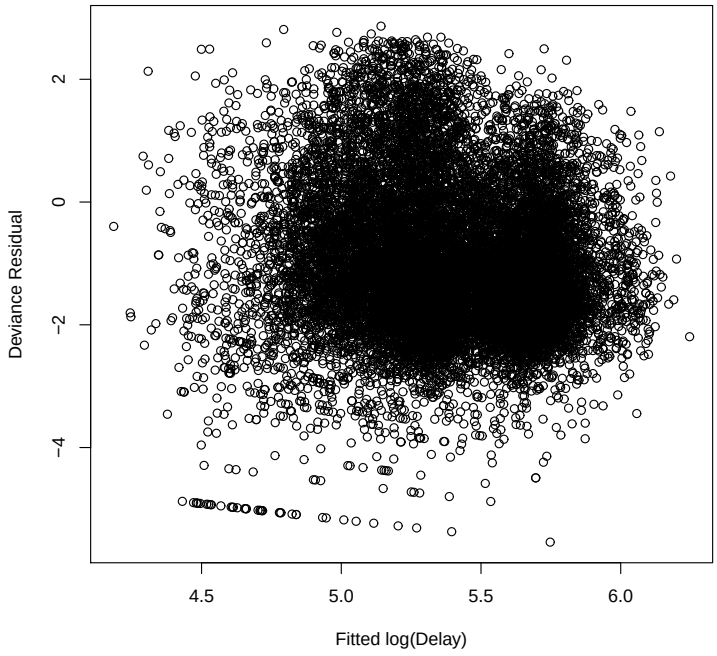
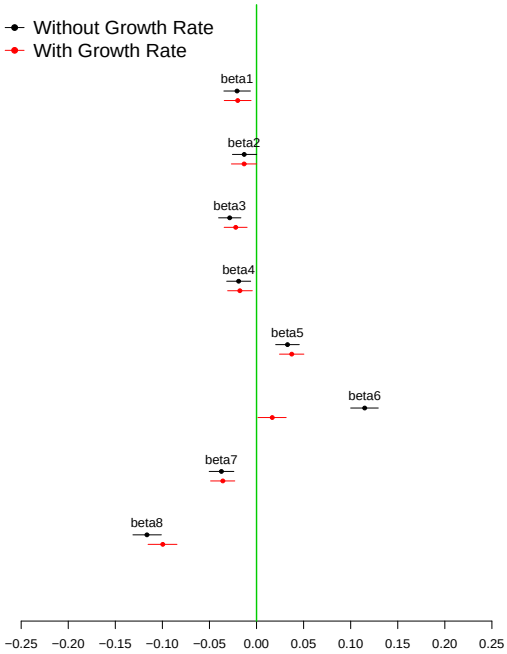
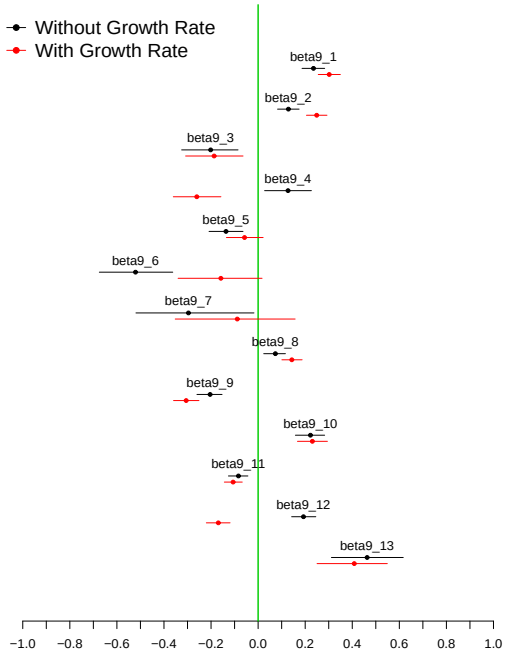


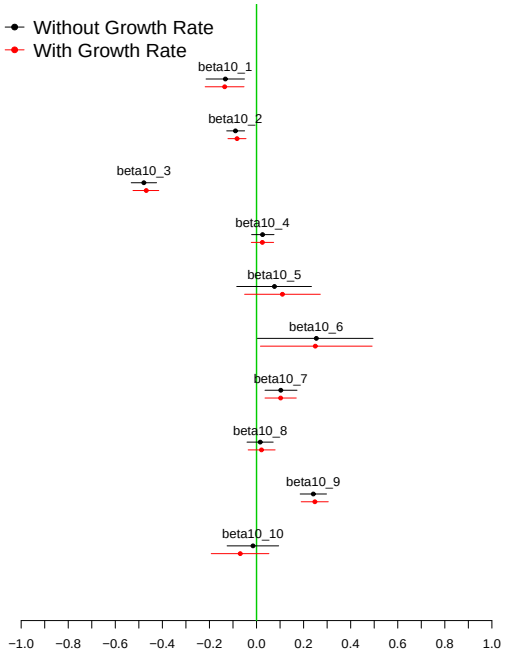
Figure 4.1: Deviance residuals of the Burr model with growth rate.



(a) $\beta_1 - \beta_8$



(b) β_9



(c) β_{10}

Figure 4.2: Comparison of the posterior estimates of the coefficients under the Burr model with and without growth rate.

The LN model

Taking the business growth into account, the probability model under the lognormal model becomes,

$$\begin{aligned} D_i &\sim LN(\mu_i, \sigma^2_{w_i}) \\ \sigma^2_{w_i} &= \sigma^2/GF_i \\ \mu_i = \eta_i &= \beta_0 + \sum_{j=1}^{8} \beta_j z_{ij} + \beta_{9,Office_i} + \beta_{10,Cause_i} \end{aligned} \tag{4.5}$$

for $i = 1, \ldots, 15860$. The β coefficients have the same prior distributions given in (4.3). For σ^2 we use the prior

$$\sigma^2 \sim IGa(0.01, 0.01).$$

The likelihood can be expressed as

$$f_w(\boldsymbol{D}|\sigma^2, \boldsymbol{\beta}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \prod_i^n \frac{1}{\sqrt{\sigma^2_{w_i}} d_i} \exp\left(-\frac{\sum (\log(d_i) - \eta_i)^2}{2\sigma^2_{w_i}}\right)$$

and the joint posterior density is

$$p_w(\sigma^2, \boldsymbol{\beta}|\boldsymbol{D}) \propto f_w(\boldsymbol{D}|\sigma^2, \boldsymbol{\beta})\pi(\sigma^2)\pi(\boldsymbol{\beta}).$$

ML estimates are obtained by maximising the likelihood given above whereas Bayesian estimates are acquired after 34000 iterations where the first 4000 iterations are considered as a burn-in process. Parameter estimates are given in Table 4.3. Similarly to the other models, the ML estimates are very close to the Bayesian estimates. The coefficients of the office are affected under the LN model when the growth is taken into account (see Figure 4.3). However, the change of this variable on the delay distribution is not as significant as in the Burr model.

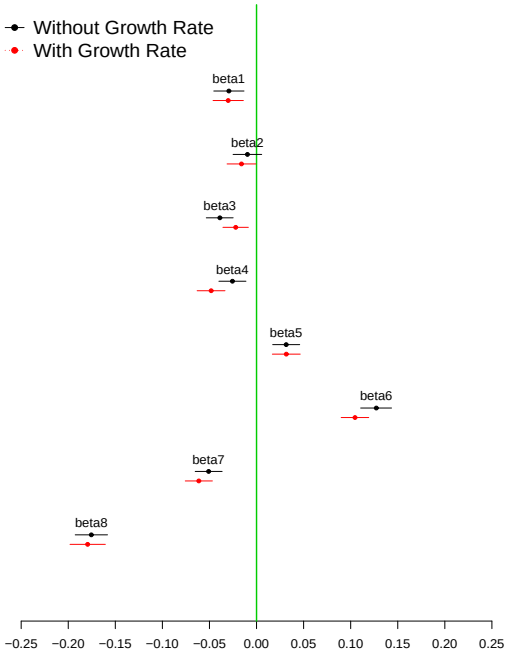
This can also be said for the coefficient of the settlement year (β_6). It is not affected from the inclusion of the growth rate in the model as much as it was under the Burr model (see Figure 4.4). The reason is that the growth rates are mostly affecting the

earlier claims and the longer delays are associated with the earlier claims. Therefore, while they are modelled better under the Burr distribution (as it is quite flexible in modelling the tail), the shorter tail of the LN distribution is not sufficient to model them.

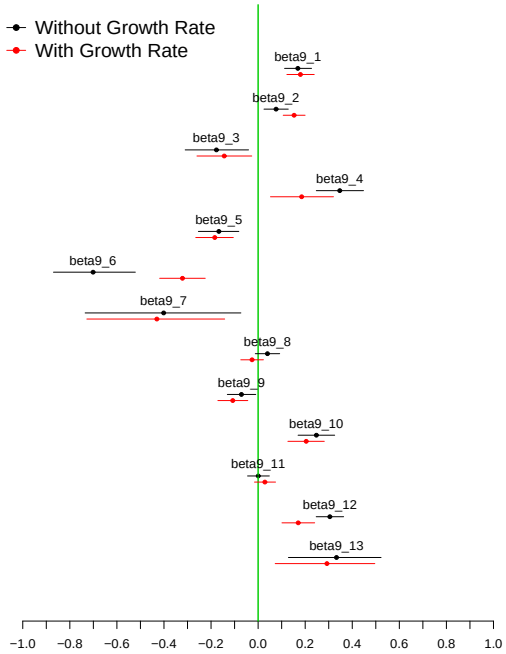
Offices 4 and 12 ($\beta_{9,4}$ and $\beta_{9,12}$) affect the delay distribution in opposite ways under the Burr and LN model. The same conclusion can be derived for Office 11 ($\beta_{9,11}$) and for TPD ($\beta_{10,10}$). However, in these two cases, the 95% credible interval suggests that the effects of these variables on the delay are not significant under the LN model and Burr model, respectively.

Table 4.3: Coefficients of the LN model with growth rate.

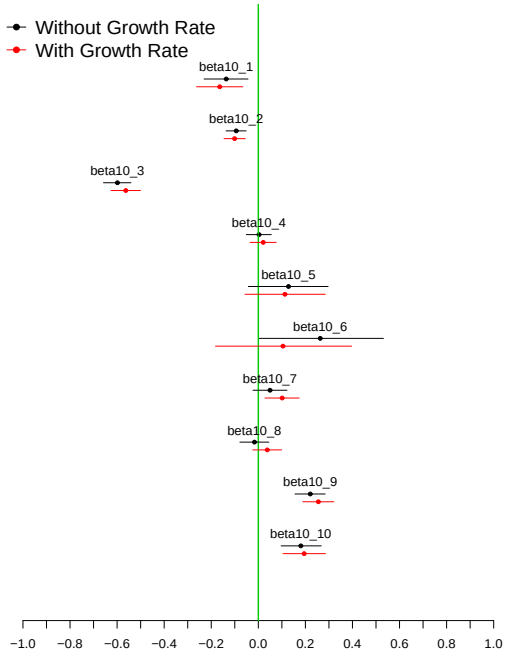
	MCMC					**MLE**	
Parameter	Mean	SD	2.5%	50%	97.5%	Mean	SD
β_0	4.7990	0.0273	4.7450	4.7990	4.8510	4.7963	0.0271
β_1	-0.0301	0.0082	-0.0462	-0.0301	-0.0140	-0.0302	0.0082
β_2	-0.0160	0.0078	-0.0313	-0.0160	-0.0007	-0.0159	0.0078
β_3	-0.0221	0.0069	-0.0357	-0.0220	-0.0087	-0.0221	0.0069
β_4	-0.0482	0.0076	-0.0631	-0.0483	-0.0335	-0.0482	0.0076
β_5	0.0316	0.0075	0.0168	0.0316	0.0464	0.0316	0.0075
β_6	0.1046	0.0075	0.0899	0.1046	0.1192	0.1048	0.0075
β_7	-0.0613	0.0074	-0.0757	-0.0613	-0.0469	-0.0613	0.0074
β_8	-0.1794	0.0096	-0.1981	-0.1795	-0.1606	-0.1796	0.0096
$\beta_{9,1}$	0.1798	0.0295	0.1222	0.1795	0.2384	0.1813	0.0295
$\beta_{9,2}$	0.1528	0.0236	0.1066	0.1526	0.1996	0.1545	0.0238
$\beta_{9,3}$	-0.1440	0.0594	-0.2599	-0.1442	-0.0270	-0.1440	0.0593
$\beta_{9,4}$	0.1850	0.0686	0.0525	0.1842	0.3201	0.1863	0.0699
$\beta_{9,5}$	-0.1842	0.0409	-0.2650	-0.1840	-0.1054	-0.1823	0.0408
$\beta_{9,6}$	-0.3214	0.0495	-0.4183	-0.3220	-0.2247	-0.3192	0.0492
$\beta_{9,7}$	-0.4302	0.1497	-0.7277	-0.4319	-0.1421	-0.4481	0.1468
$\beta_{9,8}$	-0.0258	0.0245	-0.0737	-0.0258	0.0224	-0.0243	0.0246
$\beta_{9,9}$	-0.1077	0.0323	-0.1710	-0.1079	-0.0439	-0.1059	0.0320
$\beta_{9,10}$	0.2043	0.0395	0.1265	0.2041	0.2812	0.2062	0.0392
$\beta_{9,11}$	0.0288	0.0227	-0.0154	0.0287	0.0740	0.0306	0.0226
$\beta_{9,12}$	0.1703	0.0356	0.1013	0.1702	0.2403	0.1726	0.0351
$\beta_{9,13}$	0.2924	0.1062	0.0726	0.2968	0.4958	0.2923	0.1087
$\beta_{10,1}$	-0.1639	0.0503	-0.2626	-0.1635	-0.0658	-0.1632	0.0503
$\beta_{10,2}$	-0.1007	0.0231	-0.1456	-0.1007	-0.0556	-0.0995	0.0233
$\beta_{10,3}$	-0.5631	0.0318	-0.6257	-0.5631	-0.5007	-0.5623	0.0320
$\beta_{10,4}$	0.0209	0.0284	-0.0352	0.0211	0.0760	0.0219	0.0286
$\beta_{10,5}$	0.1131	0.0874	-0.0570	0.1137	0.2848	0.1119	0.0870
$\beta_{10,6}$	0.1047	0.1475	-0.1821	0.1037	0.3964	0.0981	0.1509
$\beta_{10,7}$	0.1012	0.0371	0.0282	0.1014	0.1737	0.1023	0.0370
$\beta_{10,8}$	0.0379	0.0317	-0.0238	0.0378	0.0991	0.0395	0.0318
$\beta_{10,9}$	0.2551	0.0337	0.1890	0.2551	0.3211	0.2560	0.0336
$\beta_{10,10}$	0.1947	0.0459	0.1050	0.1945	0.2854	0.1953	0.0463
σ^2	1.0100	0.0114	0.9878	1.0100	1.0320	1.0097	



(a) $\beta_1 - \beta_8$

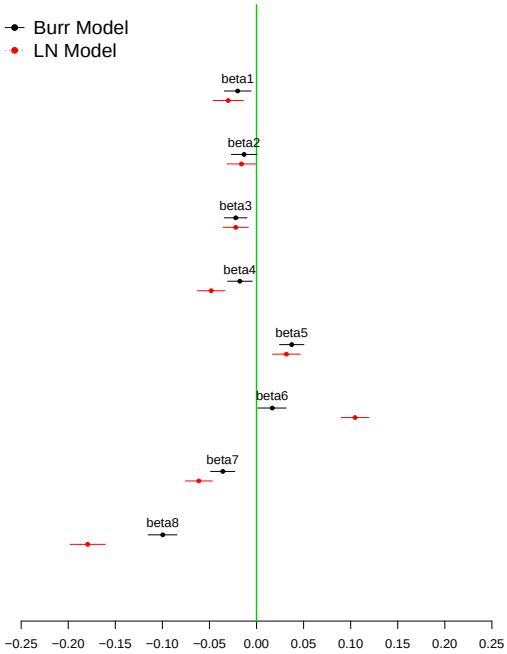


(b) β_9

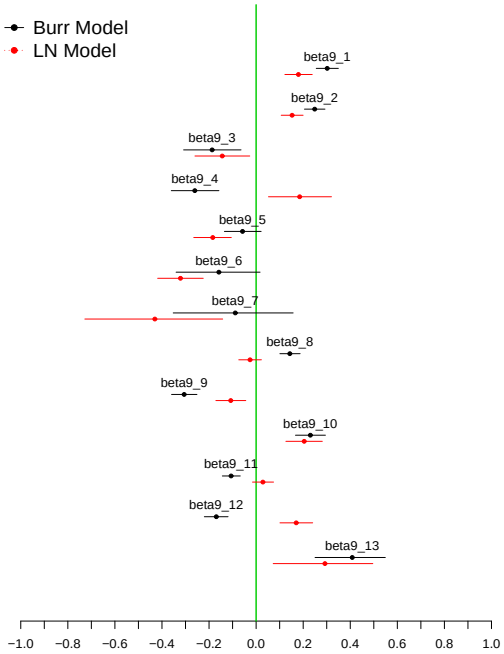


(c) β_{10}

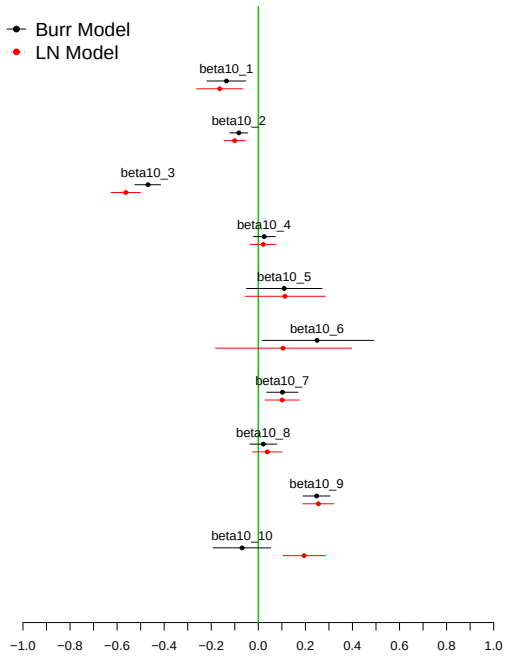
Figure 4.3: Comparison of posterior estimates of the coefficients under the LN model with and without growth rate.



(a) $\beta_1 - \beta_8$



(b) β_9



(c) β_{10}

Figure 4.4: Comparison of posterior estimates of the coefficients under the Burr and LN models.

A plot of deviance residuals against the logarithms of the fitted delays is given in Figure 4.5. A total of 3.9% of the residuals lie outside the interval [-2,2].

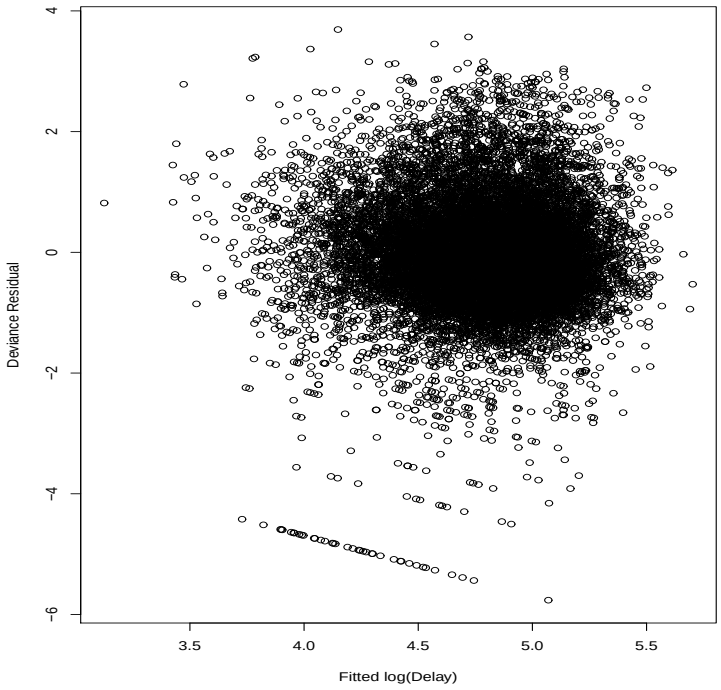


Figure 4.5: Deviance residuals of the LN model without missing values and with growth rate.

Comparison of Burr and LN models

To compare the Burr and LN models the corresponding estimates of DIC (from MCMC) and BIC (from MLE) are given in Table 4.4. Again, the Burr model has a better fit than the LN model when the growth rate is included in the models.

Table 4.4: Values of information criteria of the models with growth rate.

	$\bar{D} = -2\widehat{\log L}(\boldsymbol{\theta})$	$\hat{D} = -2\log L(\hat{\boldsymbol{\theta}})$	p_D	**DIC**	**BIC**
Burr	191228.3	191196.5	31.9	191260.2	191505.7
LN	194324.4	194293.4	31.0	194355.4	194593.2

4.3 Prediction

Table 4.5 shows the posterior estimates of the mean delay for each of the scenarios described in Table 3.7 under the Bayesian estimation procedure with the Burr and the LN model. The posterior estimates of the means and their 95% credible intervals are compared in Figure 4.6. According to these, the mean delay of a typical scenario (Scenario 1) is 178 days under the Burr model and 167 days under the LN model (these are shown with vertical lines in Figure 4.6). Since the typical scenario's 95% credible interval (169-187 days) substantially overlaps with the credible intervals of Scenarios 2 (174-192 days), 4 (162-181 days), 6 (176-201 days) & 7 (161-181 days) and 8 (163-184 days) & 9 (174-193 days), it can be said that their estimated mean delays are not significantly different from the mean of the typical scenario under the Burr model. These scenarios correspond to changing sex, smoker status, age and settlement year in the typical case. The importance of these covariates will be explored in Chapter 5. For the LN model, this is true for only Scenario 2, i.e. the estimated mean of the typical scenario of 167 days lies within the 95% credible interval of Scenario 2. On the other hand, changing office to Office 10 (Scenario 15) increases the mean delay to 249 days (199 days under the LN model) whereas death claims are settled only in 121 days (105 days under the LN model). For the policies which are in effect for 10 years (Scenario 13), the estimated mean delay is 141 days with the Burr model and 110 days with the LN model.

Table 4.5: Posterior estimates of mean delay for the scenarios in Table 3.7 for the Burr and LN model with growth rate (days).

	Burr Model					**LN Model**				
	Mean	SD	2.5%	50%	97.5%	Mean	SD	2.5%	50%	97.5%
Mean.Scn1	178.0	4.7	169.3	177.8	187.4	166.6	3.5	160.0	166.6	173.5
Mean.Scn2	182.8	4.8	173.9	182.7	192.4	172.1	3.4	165.5	172.1	178.9
Mean.Scn3	166.2	5.4	156.3	166.1	177.4	155.6	4.5	147.0	155.6	164.6
Mean.Scn4	171.0	4.9	162.1	170.9	180.8	149.3	3.6	142.4	149.3	156.6
Mean.Scn5	191.9	5.1	182.2	191.8	202.2	177.5	3.9	170.2	177.4	185.2
Mean.Scn6	187.7	6.4	175.5	187.4	200.8	180.4	5.6	169.7	180.3	191.7
Mean.Scn7	170.7	5.0	161.3	170.6	180.6	156.4	4.1	148.6	156.3	164.6
Mean.Scn8	173.1	5.3	163.3	172.9	184.1	139.8	3.6	132.9	139.7	146.9
Mean.Scn9	183.1	5.0	173.7	183.0	193.2	198.7	4.6	189.7	198.6	207.8
Mean.Scn10	182.5	4.8	173.7	182.4	192.4	173.9	3.7	166.7	173.9	181.4
Mean.Scn11	157.1	5.6	146.5	157.0	168.2	134.4	4.5	125.8	134.3	143.3
Mean.Scn12	199.8	5.2	189.9	199.6	210.4	205.1	4.4	196.7	205.1	213.8
Mean.Scn13	141.4	4.9	132.2	141.2	151.2	110.0	3.7	102.9	109.9	117.5
Mean.Scn14	169.8	17.0	139.2	168.6	205.6	117.5	6.4	105.7	117.3	130.6
Mean.Scn15	249.4	9.9	230.9	249.2	269.5	198.7	8.3	183.0	198.5	215.7
Mean.Scn16	121.1	3.9	113.5	121.0	128.8	105.0	3.3	98.6	104.9	111.5
Mean.Scn17	180.8	12.0	158.5	180.4	205.2	224.1	10.8	203.9	223.8	246.2

Changes in the means of the scenarios when the growth rate is included or excluded can be seen in Figure 4.7 for the Burr and the LN model. Under both models, posterior estimates of the means of the scenarios are higher when the growth rate is taken into account in the modelling. The significant difference between the means of Scenarios 8 & 9 and Scenario 1 disappears when growth is allowed in the model, under the Burr distribution. These scenarios are related to the change in the settlement year. Here, changing the settlement year to 1999 increases the estimated mean of Scenario 8 from 123 days to 173 days while for the most recent settlement year 2005 in Scenario 9, the estimated mean is almost the same (181 days - 183 days). This is because in this model we allow for business growth between years. By introducing weights to the model by years, we are reducing the effect of the year on the delay as this effect can partly be explained by the growth rate with this model. The significant increase in the mean delay of Scenario 8 indicates that there is a positive growth for Office 11 which we already showed in Table 4.1. The weight assigned to the claims diagnosed in 1999 for this office is around 1.75. Changing office to Office 6 (Scenario 14) can be shown as another example of a big change. Introducing the growth rate to the Burr model increases the estimated mean delay from 95 days to 170 days for this scenario. The weight assigned to this office is approximately 5, meaning that office is growing rapidly between successive years.

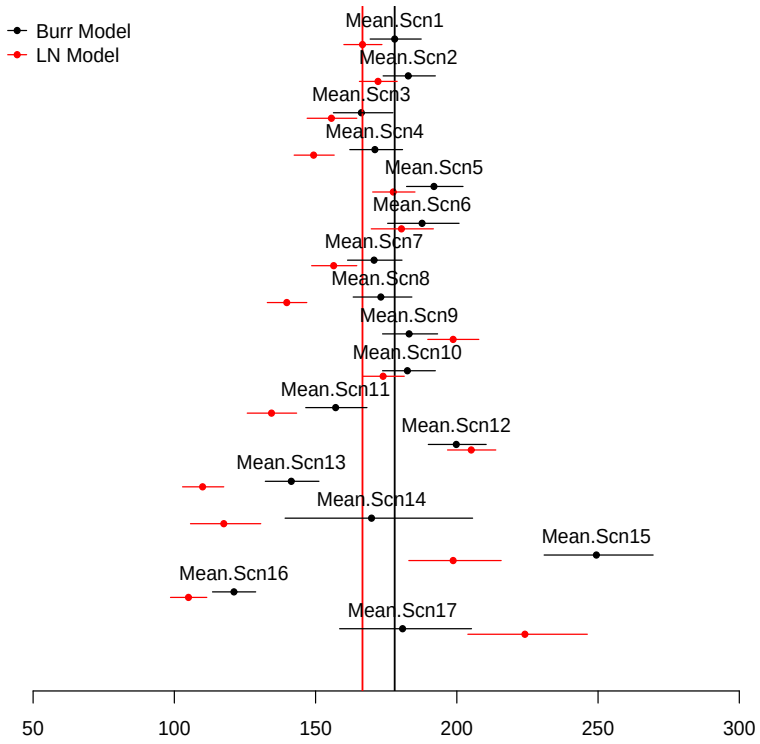
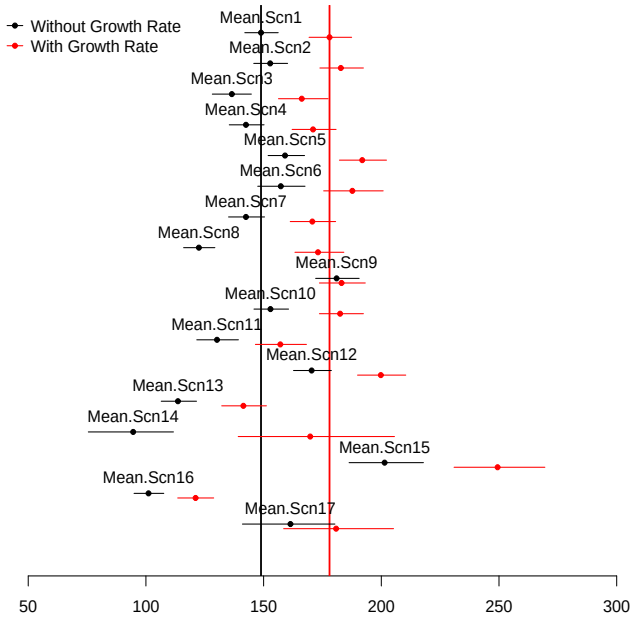
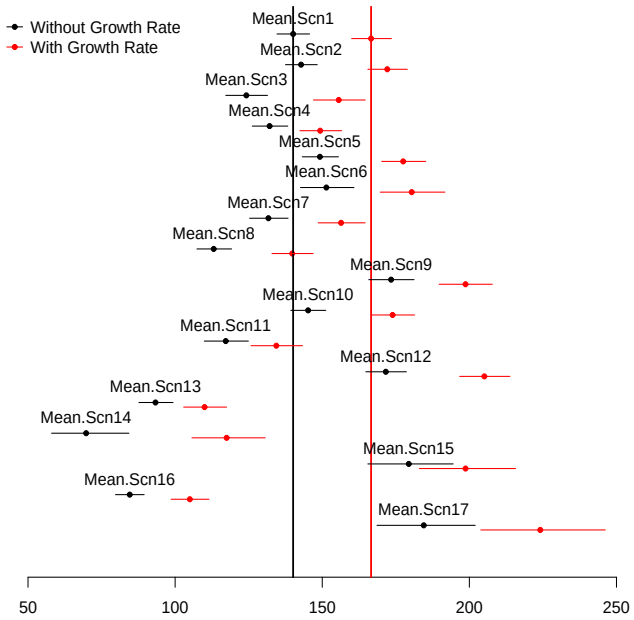


Figure 4.6: Posterior estimates of the means (days) of the scenarios under the Burr (black line) and LN model (red line) with growth rate. Bars show 95% credible intervals and bullets show posterior means. Vertical lines show the posterior means of the first scenarios under the two models.

Although the magnitudes of the mean scenarios are changed under the LN model, there is no such dramatic change between the typical scenarios and the others when we include the growth rate in the model. The difference between the typical scenario is 27 days and for most of the scenarios the difference changes between 20 days and 34 days. Some exceptions are Scenarios 4, 11 and 13 with 17 days difference and Scenarios 14 and 17 with 48 days and 40 days difference between the mean delays excluding and including the growth rate in the analysis. As it is mentioned before, the LN model is not as flexible as the Burr model in modelling the tail of the claim distribution. Since the growth factor affects the longer delays more, this can not be captured by the LN model.



(a) Burr Model



(b) LN Model

Figure 4.7: Posterior estimates of the means (days) of the scenarios under the Burr and LN model with (red line) and without (black line) growth rate. Bars show 95% credible intervals and bullets show posterior means. Vertical lines show the posterior means of the first scenarios under the two models.

Chapter 5

Selection of claim-specific covariates

5.1 Introduction

In this chapter we perform variable selection among the available claim-related factors, in order to obtain the most suitable model for describing and predicting the delay between diagnosis and settlement of claims. Different prior distribution settings for the model parameters are considered under the Bayesian approach. Here, we use the methodology introduced by Dellaportas *et al.* (2002), Ntzoufras (2002, 2009) and Ntzoufras *et al.* (2003). In particular, Gibbs variable selection methods are considered in this part. Variable selection is also carried out with classical analysis and results are compared.

In Section 5.2 claim specific covariates are selected in the absence of growth rate. Gibbs variable selection (GVS) with three different prior sets is used and results are confirmed using exact marginal likelihood findings and related Laplace approximations. For the classical analysis, a maximum likelihood based method, namely, stepwise regression, is employed. Growth rates are taken into account in Section 5.3. To find a parsimonious model, a GVS method with two different sets of priors and stepwise regression is used. For comparison purposes, a lognormal model is also fitted and model assessment measures are provided.

5.2 Selection of claim-specific covariates without growth factor

To explain the delay between dates of diagnosis and settlement, we use $p = 10$ covariates related to claim or policyholder characteristics resulting in $2^{10} = 1024$ possible models when subsets of these covariates are used. For m_t being the t^{th} model ($t = 1, \ldots, 1024$), given the marginal likelihood under m_t, $f(\boldsymbol{D}|m_t)$, and the prior model probability, $\pi(m_t)$, the posterior probability $p(m_t|\boldsymbol{D})$ can be calculated. From Bayes theorem

$$p(m_t|\boldsymbol{D}) = \frac{f(\boldsymbol{D}|m_t)\pi(m_t)}{\sum_{t=1}^{1024} f(\boldsymbol{D}|m_t)\pi(m_t)}.$$

Therefore two possible models can be compared using the so-called posterior-odds ratio

$$PO = \frac{p(m_j|\boldsymbol{D})}{p(m_k|\boldsymbol{D})} = \frac{f(\boldsymbol{D}|m_j)}{f(\boldsymbol{D}|m_k)}\frac{\pi(m_j)}{\pi(m_k)}, \qquad j \neq k \text{ with } j, k = 1, \ldots, 1024$$

where

$$B_{jk} = \frac{f(\boldsymbol{D}|m_j)}{f(\boldsymbol{D}|m_k)}$$

is the Bayes factor and $\pi(m_j)/\pi(m_k)$ is the prior odds ratio (Kass and Raftery, 1995).

The Bayes factor is given as the ratio of two marginal likelihoods. These marginal likelihoods can be obtained as

$$f(\boldsymbol{D}|m_t) = \int f(\boldsymbol{D}|\boldsymbol{\theta}_t, m_t)\pi(\boldsymbol{\theta}_t|m_t)d\boldsymbol{\theta}_t,$$

where $\boldsymbol{\theta}_t$ is the parameter vector in model m_t, $f(\boldsymbol{D}|\boldsymbol{\theta}_t, m_t)$ is the likelihood function under model m_t and $\pi(\boldsymbol{\theta}_t|m_t)$ is the prior density of the parameter vector $\boldsymbol{\theta}_t$ (Kass and Raftery, 1995).

5.2.1 Gibbs variable selection (GVS)

Let $\boldsymbol{\gamma}$ be a $p \times 1$ indicator vector where $\boldsymbol{\gamma} \in \{0, 1\}^p$ and p is the number of covariates in the model. Introducing $\boldsymbol{\gamma}$ into the linear predictor for the models in (3.4) and

(3.15) we obtain

$$\eta_i^* = \beta_0 + \sum_{j=1}^{8} \gamma_j \beta_j z_{ij} + \gamma_9 \beta_{9,Office_i} + \gamma_{10} \beta_{10,Cause_i}.$$

In Bayesian model selection, the posterior model probabilities are highly dependent on the choice of prior distributions of the model parameters. As the prior dispersion increases, more parsimonious models will be suggested. This problem is known as the Lindley-Bartlett paradox (Lindley, 1957; Bartlett, 1957). Therefore parameter prior distributions must be chosen with care. We use the following priors

$$\gamma_j \sim Bernoulli(0.5)$$

$$\beta_j \sim \ N(\mu_j, \sigma_j^2) \tag{5.1}$$

where the parameters for the distribution of β_j are considered in detail in the following sections. We note that specific priors are chosen here to facilitate variable selection, but are not used to obtain posterior estimates for the model parameters, and do not represent real prior knowledge.

Independent priors

Let us first consider a mixture of independent normal prior distributions for (5.1) (Dellaportas *et al.*, 2002) given as follows:

$$\mu_j = (1 - \gamma_j)\bar{\mu}_{\beta_j}$$

$$\sigma_j^2 = \gamma_j c^2 + (1 - \gamma_j)\bar{\sigma}_{\beta_j}^2$$

where $\bar{\mu}_{\beta_j}$ and $\bar{\sigma}_{\beta_j}^2$ are posterior estimates of the mean and variance of β_j from a pilot run with the full model and c^2 is a constant .

Three different factors, $c^2 = 100, 1000$ and $c^2 = n = 15860$, are used for the prior variance to demonstrate the Lindley-Bartlett paradox. In Tables 5.1 and 5.2 it can be seen that as the prior variances increase, the inclusion probabilities of the variables, $p(\gamma|\boldsymbol{D})$, decrease and thus simpler models are suggested. When we choose

$c^2 = 100$, posterior inclusion probabilities of the variables age, sex and smoker status (γ_1, γ_2 and γ_4) are less than 0.5 for both models (See Tables 5.1(a) and 5.2(a)). Although the inclusion probabilities decrease with increasing c^2, for the Burr model the variables having less than 0.5 inclusion probability stay the same under three different c^2s. For the LN model, the inclusion probability of policy type (γ_5) also goes below 0.5 for $c^2 = 1000$ and $c^2 = n$. Note that the SD columns in the tables give the standard deviations for the posterior distributions of the inclusion probabilities.

These results are consistent with the posterior model probabilities given in Tables 5.1 and 5.2. That is, model number 1013, which excludes variables age, sex and smoker status is supported for all cases under the Burr model with decreasing posterior model probabilities. The same model has the highest probability under the lognormal model when $c^2 = 100$. However, the posterior model probabilities of the top three models are very close to each other with posterior odds close to 1. Since we use the same a-priori probabilities for each possible model, the posterior odds will be equal to the Bayes factor. The Bayes factor can be used to compare nested models. According to Kass and Raftery (1995), the difference between models is negligible when the Bayes factor is less than 3. Similarly, when $c^2 = 1000$, the posterior odds of model 997 to model 1013 is $m_2/m_1 = 2.82 < 3$ and thus they are barely different from each other. For the last case, model 997 has the highest posterior probability. This model excludes policy duration as well as age, sex and smoker status.

In this example, we saw how the Lindley-Bartlett paradox is activated when these independent priors are used. Since the posterior model probabilities are highly dependent on the choice of the prior variance, this set of priors is not always useful. Here, especially under the Burr distribution, we have consistent model selection results; this is probably because we have a very large data set.

Table 5.1: Parameter inclusion probabilities and model probabilities under the Burr model with independent normal priors.

(a) Inclusion probabilities for various prior variances.

	$c^2 = 100$		$c^2 = 1000$		$c^2 = n$	
Parameter	$p(\gamma\|\boldsymbol{D})$	SD	$p(\gamma\|\boldsymbol{D})$	SD	$p(\gamma\|\boldsymbol{D})$	SD
γ_1	0.0487	0.2152	0.0238	0.1525	0.0059	0.0766
γ_2	0.0137	0.1162	0.0032	0.0562	0.0009	0.0300
γ_3	0.9541	0.2093	0.8971	0.3038	0.6009	0.4897
γ_4	0.0570	0.2318	0.0249	0.1558	0.0059	0.0766
γ_5	0.9732	0.1614	0.9154	0.2783	0.6969	0.4596
γ_6	1	0	1	0	1	0
γ_7	0.9999	0.0082	0.9752	0.1554	0.9946	0.0731
γ_8	1	0	1	0	1	0
γ_9	1	0	1	0	1	0
γ_{10}	1	0	1	0	1	0

(b) Posterior model probabilities - $c^2 = 100$

Rank	Model No	Model	$f(m\|\boldsymbol{D})$	PO(m_{1013}/m_t)
1	m_{1013}	$x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.8266	1.00
2	m_{1021}	$x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0530	15.61
3	m_{1014}	$x_1 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0433	19.09
4	m_{1009}	$x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0313	26.38
5	m_{997}	$x_3 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0151	54.63
6	m_{1015}	$x_2 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0120	68.88
7	m_{993}	$x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0110	75.15
8	m_{1022}	$x_1 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0027	309.94
9	m_{1010}	$x_1 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0021	387.53
10	m_{1017}	$x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0008	1033.25

(c) Posterior model probabilities - $c^2 = 1000$

Rank	Model No	Model	$f(m\|\boldsymbol{D})$	PO(m_{1013}/m_t)
1	m_{1013}	$x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.8052	1.00
2	m_{1009}	$x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0493	16.34
3	m_{993}	$x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0382	21.08
4	m_{997}	$x_3 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0318	25.34
5	m_{1021}	$x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0232	34.75
6	m_{1014}	$x_1 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0201	40.06
7	m_{929}	$x_6 + x_8 + x_9 + x_{10}$	0.0089	90.81
8	m_{949}	$x_3 + x_5 + x_6 + x_8 + x_9 + x_{10}$	0.0088	91.50
9	m_{933}	$x_3 + x_6 + x_8 + x_9 + x_{10}$	0.0043	187.26
10	m_{1010}	$x_1 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0032	249.06

(d) Posterior model probabilities - $c^2 = n$

Rank	Model No	Model	$f(m\|\boldsymbol{D})$	PO(m_{1013}/m_t)
1	m_{1013}	$x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.5442	1.00
2	m_{993}	$x_6 + x_7 + x_8 + x_9 + x_{10}$	0.2476	2.20
3	m_{1009}	$x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.1422	3.83
4	m_{997}	$x_3 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0480	11.33
5	m_{1021}	$x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0051	106.02
6	m_{929}	$x_6 + x_8 + x_9 + x_{10}$	0.0045	120.05
7	m_{1014}	$x_1 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0027	199.12
8	m_{994}	$x_1 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0017	314.02
9	m_{1014}	$x_1 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0012	441.36
10	m_{945}	$x_5 + x_6 + x_8 + x_9 + x_{10}$	0.0007	816.26

Table 5.2: Parameter inclusion probabilities and model probabilities under the LN model with independent normal priors.

(a) Inclusion probabilities for various prior variances.

	$c^2 = 100$		$c^2 = 1000$		$c^2 = n$	
Parameter	$p(\gamma\|\boldsymbol{D})$	SD	$p(\gamma\|\boldsymbol{D})$	SD	$p(\gamma\|\boldsymbol{D})$	SD
γ_1	0.3292	0.4699	0.0946	0.2926	0.0181	0.1332
γ_2	0.0024	0.0493	0.0008	0.0289	0.0002	0.0129
γ_3	0.9954	0.0679	0.9781	0.1462	0.9083	0.2886
γ_4	0.2137	0.4099	0.0661	0.2485	0.0158	0.1248
γ_5	0.6331	0.4819	0.2900	0.4538	0.0779	0.2680
γ_6	1	0	1	0	1	0
γ_7	1	0	0.9999	0.0082	0.9996	0.0192
γ_8	1	0	1	0	1	0
γ_9	1	0	1	0	1	0
γ_{10}	1	0	1	0	1	0

(b) Posterior model probabilities - $c^2 = 100$

Rank	Model No	Model	$f(m\|\boldsymbol{D})$	PO(m_{1013}/m_t)
1	m_{1013}	$x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.2816	1.00
2	m_{997}	$x_3 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.2524	1.12
3	m_{1014}	$x_1 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.1999	1.41
4	m_{1021}	$x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0803	3.51
5	m_{1022}	$x_1 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0692	4.07
6	m_{1005}	$x_3 + x_4 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0521	5.40
7	m_{998}	$x_1 + x_3 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0464	6.07
8	m_{1006}	$x_1 + x_3 + x_4 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0111	25.44
9	m_{993}	$x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0022	128.00
10	m_{994}	$x_1 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0015	191.96

(c) Posterior model probabilities - $c^2 = 1000$

Rank	Model No	Model	$f(m\|\boldsymbol{D})$	PO(m_{1013}/m_t)
1	m_{997}	$x_3 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.6112	0.35
2	m_{1013}	$x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.2168	1.00
3	m_{1014}	$x_1 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0484	4.48
4	m_{1005}	$x_3 + x_4 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0390	5.56
5	m_{998}	$x_1 + x_3 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0354	6.13
6	m_{1021}	$x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0188	11.55
7	m_{993}	$x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0180	12.06
8	m_{1022}	$x_1 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0051	42.24
9	m_{994}	$x_1 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0026	82.34
10	m_{1006}	$x_1 + x_3 + x_4 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0026	83.38

(d) Posterior model probabilities - $c^2 = n$

Rank	Model No	Model	$f(m\|\boldsymbol{D})$	PO(m_{1013}/m_t)
1	m_{997}	$x_3 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.8062	0.09
2	m_{993}	$x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0871	0.83
3	m_{1013}	$x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0721	1.00
4	m_{1005}	$x_3 + x_4 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0127	5.69
5	m_{998}	$x_1 + x_3 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0115	6.25
6	m_{1014}	$x_1 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0034	21.00
7	m_{994}	$x_1 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0027	26.70
8	m_{1021}	$x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0016	46.01
9	m_{1001}	$x_4 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0013	56.91
10	m_{1009}	$x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0005	144.20

Empirical priors

We use empirical priors following Ntzoufras (2009). Here, $\beta_j \sim N(\mu_j, \sigma_j^2)$ and the prior means and variances of the parameters are as follows

$$\mu_j = \bar{\mu}_{\beta_j}, \qquad j = 1, \ldots, p$$

$$\sigma_j^2 = \{\gamma_j n + (1 - \gamma_j)\}\, \bar{\sigma}_{\beta_j}^2$$

where $\bar{\mu}_{\beta_j}$ and $\bar{\sigma}_{\beta_j}^2$ are posterior estimates of the mean and variance of β_j from a pilot run with the full model. This set of priors can be considered as empirical priors because the data we have already observed are used in the prior. However the observed data count as one additional data point only, and therefore the double usage of the data is low (see Ntzoufras (2009, page 408)).

Zellner's g-prior

A prior based on Zellner's g-prior is also considered. For a normal model, Zellner (1986) suggested a Normal-Inverse Gamma conjugate prior for model m of the form

$$\boldsymbol{\beta}_m | \sigma^2, m \sim MVN(\boldsymbol{\mu}_{\boldsymbol{\beta}_m}, c^2 (\boldsymbol{Z}'_m \boldsymbol{Z}_m)^{-1} \sigma^2) \tag{5.2}$$

where $\boldsymbol{Z}_m$ is the $n \times (p_m + 1)$ standardised design matrix, $\boldsymbol{\beta}_m$ is the $(p_m + 1) \times 1$ coefficient vector of model m with p_m being the number of covariates involved in the linear part of model m and c^2 is a constant. Defining $c^2 = n$, we can specify Zellner's g-prior for the parameter vector $\boldsymbol{\beta}_m$ using the following multivariate normal prior distribution

$$\boldsymbol{\beta}_m \sim MVN(\boldsymbol{\mu}_{\boldsymbol{\beta}_m}, \boldsymbol{S}^{-1})$$

Here, $\boldsymbol{\mu}_{\boldsymbol{\beta}_m} = \mu_0, \ldots, \mu_p$ is the prior mean vector with

$$\mu_j = (1 - \gamma_j)\bar{\mu}_{\beta_j}, \qquad j = 1, \ldots, p$$

and the (j,k) element of the $(p+1)\times(p+1)$ matrix $\boldsymbol{S}$ is given by

$$S_{j,k} = \frac{\gamma_j\gamma_k}{n\sigma^2}(\boldsymbol{Z}'\boldsymbol{Z})_{jk} + (1-\gamma_j\gamma_k)I\{j=k\}\bar{\sigma}^{-2}_{\beta_j}, \qquad j,k = 1,\ldots,p.$$

with $\gamma_0 = 1$ and $\bar{\mu}_{\beta_j}$, $\bar{\sigma}^2_{\beta_j}$ as before. For the Burr model on the other hand, (5.2) can not be used directly. However, an approximation to (5.2) can be employed using the information matrix as the unknown prior variance-covariance matrix to give

$$\boldsymbol{\beta}_m \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}_m}, n\hat{\boldsymbol{\Sigma}}),$$

where $\hat{\boldsymbol{\Sigma}} = (-D^2 l(\hat{\boldsymbol{\beta}}))^{-1}$ is the estimated variance-covariance matrix with $D^2 l(\hat{\boldsymbol{\beta}})$ being the Hessian matrix of the log-likelihood evaluated at the maximum likelihood estimate (Kass and Wasserman, 1995; Ntzoufras *et al.*, 2003).

Model selection results

Inclusion and posterior model probabilities for empirical and Zellner priors are shown in Tables 5.3 and 5.4 for Burr and LN models, respectively. Results under these two prior sets are very close to each other. This is not surprising since both of these priors considered here are 'unit information priors' which is introduced by Kass and Wasserman (1995). They are very weakly informative given the size of the data and the fact that they use information equivalent to that contained in a single data point.

Under the Burr model, the inclusion probability of sex (γ_2) is around 0.08 with both sets of priors implying that there is high probability of this variable being excluded from the best models. On the other hand, inclusion probabilities of age and smoker status (γ_1 and γ_4) are slightly under 0.5 (see Table 5.3(a)). As can be seen from Table 5.3(b) and 5.3(c), the highest posterior model probability is for model 1013 under both prior sets which excludes all these three variables.

The same sets of priors are considered for the LN model with both giving almost the same posterior model probabilities. The ten models with the highest posterior model probabilities are given in Table 5.4. The only variable which has an inclusion probability less than 0.5 is sex. The probability of this variable to be included in the model is less than 0.02. Model 1022, which excludes this variable, has the highest

posterior model probability.

Considering these results, it can be concluded that the best model under the Burr distribution is model 1013 and under the lognormal distribution it is model 1022. However the first four models under the Burr distribution, including model 1022, have posterior odds less than 2. Therefore we can conclude that the best model under the lognormal assumption is not inconsistent with the model selection under the Burr distribution.

5.2.2 Variable selection using marginal likelihoods

Marginal likelihood for LN model

For normal regression models, the exact marginal likelihoods, and thus the model probabilities, can be calculated analytically when conjugate distributions for the model parameters are used. In this case, a Normal-Inverse Gamma conjugate prior for model m, where $m \in (m_1, \ldots, m_{1024})$ is used.

The required marginal likelihood for the lognormal model has the form

$$f(\boldsymbol{D}|m) = \int f(\boldsymbol{D}|\boldsymbol{\theta}_m, m)\pi(\boldsymbol{\theta}_m|m)d\boldsymbol{\theta}_m, \tag{5.3}$$

where $\boldsymbol{\theta}_m$ is the parameter vector of model m, including σ^2 and the $(p_m + 1) \times 1$ coefficient vector $\boldsymbol{\beta}_m$. The logarithm of the delay ($\log(\boldsymbol{D})$) is distributed as

$$\log(\boldsymbol{D})|\boldsymbol{\beta}_m, \sigma^2, m \sim MVN(\boldsymbol{Z}_m\boldsymbol{\beta}_m, \sigma^2\boldsymbol{I}_n)$$

with $\boldsymbol{I}_n$ being the $n \times n$ identity matrix. In (5.3)

$$\pi(\boldsymbol{\theta}_m|m) = \pi(\boldsymbol{\beta}_m, \sigma^2|m) \propto \pi(\boldsymbol{\beta}_m|\sigma^2, m)\pi(\sigma^2)$$

with prior distributions

$$\pi(\boldsymbol{\beta}_m|\sigma^2, m) \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}_m}, c^2\boldsymbol{V}_m\sigma^2)$$

$$\pi(\sigma^2) \sim IGa(a, b)$$

Table 5.3: Parameter inclusion probabilities and model probabilities under the Burr model with empirical and Zellner's g-priors.

(a) Inclusion probabilities for empirical priors and Zellner's g-prior.

	Empirical Priors		Zellner's g-Prior	
Parameter	$p(\gamma\|\boldsymbol{D})$	SD	$p(\gamma\|\boldsymbol{D})$	SD
γ_1	0.4460	0.4971	0.4171	0.4931
γ_2	0.0843	0.2778	0.0890	0.2847
γ_3	0.9992	0.0277	0.9973	0.0519
γ_4	0.4420	0.4966	0.4113	0.4921
γ_5	1	0	1	0
γ_6	1	0	1	0
γ_7	1	0	1	0
γ_8	1	0	1	0
γ_9	1	0	1	0
γ_{10}	1	0	1	0

(b) Posterior model probabilities - Empirical Priors

Rank	Model No	Model	$f(m\|\boldsymbol{D})$	PO(m_{1013}/m_t)
1	m_{1013}	$x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.2847	1.00
2	m_{1014}	$x_1 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.2180	1.31
3	m_{1021}	$x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.2120	1.34
4	m_{1022}	$x_1 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.2002	1.42
5	m_{1015}	$x_2 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0377	7.55
6	m_{1023}	$x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0193	14.75
7	m_{1016}	$x_1 + x_2 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0171	16.62
8	m_{1024}	$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0101	28.10
9	m_{1010}	$x_1 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0003	949.00
10	m_{1018}	$x_1 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0002	1220.32

(c) Posterior model probabilities - Zellner's g-Prior

Rank	Model No	Model	$f(m\|\boldsymbol{D})$	PO(m_{1013}/m_t)
1	m_{1013}	$x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.3076	1.00
2	m_{1014}	$x_1 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.2186	1.41
3	m_{1021}	$x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.2166	1.42
4	m_{1022}	$x_1 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.1656	1.86
5	m_{1015}	$x_2 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0389	7.90
6	m_{1016}	$x_1 + x_2 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0217	14.16
7	m_{1023}	$x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0184	16.72
8	m_{1024}	$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0098	31.28
9	m_{1009}	$x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0010	318.20
10	m_{1010}	$x_1 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0008	384.50

Table 5.4: Parameter inclusion probabilities and model probabilities under the LN model with empirical and Zellner's g-priors.

(a) Inclusion probabilities for empirical priors and Zellner's g-prior.

	Empirical Priors		Zellner's g-Prior	
Parameter	$p(\gamma\|\boldsymbol{D})$	SD	$p(\gamma\|\boldsymbol{D})$	SD
γ_1	0.8870	0.3166	0.8840	0.3203
γ_2	0.0185	0.1349	0.0179	0.1327
γ_3	0.9997	0.0173	0.9999	0.0100
γ_4	0.7876	0.4090	0.7891	0.4079
γ_5	0.9785	0.1449	0.9796	0.1415
γ_6	1	0	1	0
γ_7	1	0	1	0
γ_8	1	0	1	0
γ_9	1	0	1	0
γ_{10}	1	0	1	0

(b) Posterior model probabilities - Empirical Priors

Rank	Model No	Model	$f(m\|\boldsymbol{D})$	PO(m_{1022}/m_t)
1	m_{1022}	$x_1 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.6808	1.00
2	m_{1014}	$x_1 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.1760	3.87
3	m_{1021}	$x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0780	8.72
4	m_{1013}	$x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0254	26.83
5	m_{1024}	$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0120	56.88
6	m_{1006}	$x_1 + x_3 + x_4 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0096	70.67
7	m_{1005}	$x_3 + x_4 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0047	143.84
8	m_{998}	$x_1 + x_3 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0045	152.41
9	m_{1016}	$x_1 + x_2 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0036	189.11
10	m_{997}	$x_3 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0021	319.17

(c) Posterior model probabilities - Zellner's g-Prior

Rank	Model No	Model	$f(m\|\boldsymbol{D})$	PO(m_{1022}/m_t)
1	m_{1022}	$x_1 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.6828	1.00
2	m_{1014}	$x_1 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.1737	3.93
3	m_{1021}	$x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0787	8.68
4	m_{1013}	$x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0269	25.35
5	m_{1024}	$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0110	62.07
6	m_{1006}	$x_1 + x_3 + x_4 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0092	74.48
7	m_{1005}	$x_3 + x_4 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0049	140.29
8	m_{998}	$x_1 + x_3 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0037	182.91
9	m_{1016}	$x_1 + x_2 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0033	209.00
10	m_{1023}	$x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0022	315.09

where $\boldsymbol{V}_m$ is the prior variance-covariance matrix, obtained from the pilot run. Then, following Ntzoufras (2009), the exact marginal log-likelihood can be expressed as

$$\begin{aligned}\log(f(\boldsymbol{D}|m)) = &- p_m \log(c) + 0.5\left(\log(|\tilde{\boldsymbol{\Sigma}}_m|) - \log(|\boldsymbol{V}_m|)\right) - \\ &(0.5n + a)\log\left(0.5\left((\log(\boldsymbol{D}))' \log(\boldsymbol{D}) - \tilde{\boldsymbol{\beta}}_m' \tilde{\Sigma}_m^{-1} \tilde{\boldsymbol{\beta}}_m + c^{-2} \boldsymbol{\mu}_{\boldsymbol{\beta}_m}' \boldsymbol{V}_m^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}_m}\right) + b\right) \\ &+ \kappa - \sum_i^n \log(D_i)\end{aligned} \tag{5.4}$$

where $\tilde{\boldsymbol{\Sigma}}$ is the posterior variance-covariance matrix given by

$$\tilde{\boldsymbol{\Sigma}}_m = \left(\boldsymbol{Z}'_m \boldsymbol{Z}_m + c^{-2}\boldsymbol{V}_m^{-1}\right)^{-1}$$

and the posterior mean $\tilde{\boldsymbol{\beta}}$ is written as

$$\tilde{\boldsymbol{\beta}}_m = \tilde{\boldsymbol{\Sigma}}_m \left(\boldsymbol{Z}'_m \boldsymbol{Z}_m \hat{\boldsymbol{\beta}} + c^{-2}\boldsymbol{V}_m^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}_m}\right)$$

with $\kappa = a\log(b) - \log(\Gamma(a)) - 0.5n\log(2\pi) + \log(\Gamma(0.5n + a))$. Here $\hat{\boldsymbol{\beta}}$ denotes the MLE of the coefficients.

Table 5.5 gives the ten highest marginal probability models using empirical priors. Comparing this table with Table 5.4(c), it can be seen that the order of the first seven models is the same, with very similar posterior model probabilities. Here, we also conclude that the best model under the LN model is model 1022 which excludes age.

Table 5.5: Exact marginal likelihoods (EML) for the lognormal model.

Rank	Model No	Model	Marginal Loglik	$p(m\|\boldsymbol{D})$	PO(m_{1022}/m_t)
1	m_{1022}	$x_1 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	-96214.18	0.6815	1.00
2	m_{1014}	$x_1 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	-96215.54	0.1749	3.90
3	m_{1021}	$x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	-96216.34	0.0786	8.67
4	m_{1013}	$x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	-96217.45	0.0259	26.31
5	m_{1024}	$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	-96218.24	0.0118	57.97
6	m_{1006}	$x_1 + x_3 + x_4 + x_6 + x_7 + x_8 + x_9 + x_{10}$	-96218.43	0.0097	70.11
7	m_{1005}	$x_3 + x_4 + x_6 + x_7 + x_8 + x_9 + x_{10}$	-96219.22	0.0044	154.47
8	m_{1016}	$x_1 + x_2 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	-96219.35	0.0039	175.91
9	m_{998}	$x_1 + x_3 + x_6 + x_7 + x_8 + x_9 + x_{10}$	-96219.36	0.0038	177.68
10	m_{997}	$x_3 + x_6 + x_7 + x_8 + x_9 + x_{10}$	-96219.96	0.0021	323.76

Laplace approximation for the Burr model

Analytic computation of (5.3) may not be tractable for more complicated models. In these cases, approximations such as Laplace's method can be used to obtain an adequate approximation for marginal likelihoods. Since this method relies on an underlying normal distribution, the approximations will be more effective for symmetric posterior distributions. However, in practice, for large sample sizes this method gives reasonable approximations to marginal likelihoods (Gelman *et al.*, 2000). A detailed description of the method can be found in Kass and Raftery (1995). Here we only give the resulting approximation, i.e.

$$\hat{f}(\boldsymbol{D}|m) = (2\pi)^{d_m/2}|\tilde{\boldsymbol{\Sigma}}_m|^{1/2} f(\boldsymbol{D}|\tilde{\boldsymbol{\theta}}_m, m)\pi(\tilde{\boldsymbol{\theta}}_m|m)$$

where $\tilde{\boldsymbol{\theta}}_m$ is the vector of the posterior modes of the parameters under model m with dimension d_m, and $\tilde{\boldsymbol{\Sigma}} = (-D^2 l(\tilde{\boldsymbol{\beta}}))^{-1}$ is the covariance matrix where $D^2 l(\tilde{\boldsymbol{\beta}})$ is the Hessian matrix of the likelihood evaluated at the posterior modes of the parameters. However, obtaining the covariance matrix at the posterior modes is not computationally easy. Thus, following Kass and Raftery (1995), we calculate the marginal likelihoods with a variant of the Laplace approximation, using the covariance matrix $\hat{\boldsymbol{\Sigma}}_m$ which employs MLE estimates of the parameters to calculate the Hessian matrix. That is

$$\hat{f}_{MLE}(\boldsymbol{D}|m) = (2\pi)^{d_m/2}|\hat{\boldsymbol{\Sigma}}_m|^{1/2} f(\boldsymbol{D}|\hat{\boldsymbol{\theta}}_m, m)\pi(\hat{\boldsymbol{\theta}}_m|m).$$

The approximated marginal likelihoods, $\hat{f}_{MLE}(\boldsymbol{D}|m)$, for the Burr model are presented in Table 5.6, together with the resulting posterior model probabilities $p(m|\boldsymbol{D})$ and posterior odds.

The results coincide with those obtained using the GVS method. Model 1013 has again the highest approximated likelihood under both prior sets. When Tables 5.3 and 5.6 are compared it can be seen that the first five models are always the same. The ordering of the second and the third model in the Laplace approximation using empirical priors changes but the difference between the likelihoods is very small.

Table 5.6: Laplace Approximation for the Burr Model.

(a) Approximation with empirical priors.

Rank	Model No	Model	$\hat{f}_{MLE}(\boldsymbol{D}\|m)$	$p(m\|\boldsymbol{D})$	PO(m_{1013}/m_t)
1	m_{1013}	$x_3+x_5+x_6+x_7+x_8+x_9+x_{10}$	-95319.67	0.2892	1.00
2	m_{1021}	$x_3+x_4+x_5+x_6+x_7+x_8+x_9+x_{10}$	-95319.94	0.2199	1.32
3	m_{1014}	$x_1+x_3+x_5+x_6+x_7+x_8+x_9+x_{10}$	-95319.97	0.2148	1.35
4	m_{1022}	$x_1+x_3+x_4+x_5+x_6+x_7+x_8+x_9+x_{10}$	-95320.08	0.1924	1.50
5	m_{1015}	$x_2+x_3+x_5+x_6+x_7+x_8+x_9+x_{10}$	-95321.74	0.0364	7.94
6	m_{1023}	$x_2+x_3+x_4+x_5+x_6+x_7+x_8+x_9+x_{10}$	-95322.40	0.0188	15.36
7	m_{1016}	$x_1+x_2+x_3+x_5+x_6+x_7+x_8+x_9+x_{10}$	-95322.56	0.0160	18.05
8	m_{1024}	$x_1+x_2+x_3+x_4+x_5+x_6+x_7+x_8+x_9+x_{10}$	-95323.03	0.0100	28.96
9	m_{1010}	$x_1+x_5+x_6+x_7+x_8+x_9+x_{10}$	-95325.72	0.0007	424.97
10	m_{1009}	$x_5+x_6+x_7+x_8+x_9+x_{10}$	-95326.08	0.0005	610.07

(b) Approximation with Zellner's g-prior.

Rank	Model No	Model	$\hat{f}_{MLE}(\boldsymbol{D}\|m)$	$p(m\|\boldsymbol{D})$	PO(m_{1013}/m_t)
1	m_{1013}	$x_3+x_5+x_6+x_7+x_8+x_9+x_{10}$	-95320.82	0.2847	1.00
2	m_{1014}	$x_1+x_3+x_5+x_6+x_7+x_8+x_9+x_{10}$	-95321.09	0.2182	1.30
3	m_{1021}	$x_3+x_4+x_5+x_6+x_7+x_8+x_9+x_{10}$	-95321.09	0.2178	1.31
4	m_{1022}	$x_1+x_3+x_4+x_5+x_6+x_7+x_8+x_9+x_{10}$	-95321.19	0.1967	1.45
5	m_{1015}	$x_2+x_3+x_5+x_6+x_7+x_8+x_9+x_{10}$	-95322.91	0.0355	8.03
6	m_{1023}	$x_2+x_3+x_4+x_5+x_6+x_7+x_8+x_9+x_{10}$	-95323.56	0.0184	15.44
7	m_{1016}	$x_1+x_2+x_3+x_5+x_6+x_7+x_8+x_9+x_{10}$	-95323.69	0.0161	17.68
8	m_{1024}	$x_1+x_2+x_3+x_4+x_5+x_6+x_7+x_8+x_9+x_{10}$	-95324.16	0.0101	28.20
9	m_{1010}	$x_1+x_5+x_6+x_7+x_8+x_9+x_{10}$	-95326.82	0.0007	404.12
10	m_{1009}	$x_5+x_6+x_7+x_8+x_9+x_{10}$	-95327.22	0.0005	598.82

5.2.3 Maximum likelihood based methods

Backward Stepwise Method: This method is based on the likelihood ratio test (LRT). LRT $= -2(l_0 - l_1)$ where l_0 and l_1 denote the log-likelihood functions of the null and the alternative model with df_0 and df_1 degrees of freedom, respectively. For large number of observations (n), the distribution of LRT approaches the χ^2 distribution with $\text{df}_0 - \text{df}_1$ degrees of freedom (Miller and Miller, 2004). For the null hypothesis $H_0 : x_p = 0$ where x_p has 1 degree of freedom

$$\Delta BIC = -2(l_0 - l_1) - \log(n)$$

$$= LRT - \log(n).$$

The null hypothesis will not be rejected when $LRT \leq \log(n)$ at some significance level α. Here, $\log(n) \approx 9.7$ with n being 15860 and $\chi^2_1 = 9.7$ at $\alpha \approx 0.002$. This significance level seems reasonable considering the size of the data set.

We note here that the forward stepwise method gives the same best models under the Burr and LN distributions. Therefore we present only the results obtained by the backward method.

Tables 5.7 and 5.8 show log-likelihood and BIC values of different models under the Burr and lognormal distributions, respectively. The LRT results for each step are summarised in Table 5.9. Since this is a backward method, we start with the full model, which is model 1024, and drop one variable at each step, unless there are no variables to drop (reject the null hypothesis). For example, for the Burr distribution (Table 5.7), the first step is to drop each one of the 10 covariates from the model at a time and compare these 10 models with the full model, model 1024. Sorting these models according to their BIC values, the model which excludes x_2 (sex) has the smallest BIC with LRT giving 3.8 for the hypothesis of $x_2 = 0$ (see Table 5.9). Since this is smaller than the critical value of 9.7, we conclude to drop this variable from the model and continue to the second step. For this step our full model is model 1022, which excludes variable x_2. From this model, we drop each of the remaining 9 variables (one at a time) and find their BIC values. Sorting them in ascending order suggests that the model, 1014, which does not include variable x_4 (smoker status) as well as x_2 is better than model 1022 (LRT $= 9.3 < 9.7$). Continuing to the third step, we find that dropping x_1 (age) from model 1014, gives a better model (model 1013) as LRT $= 9.3 < 9.7$ for the hypothesis of $x_1 = 0$. On the fourth step, none of the models excluding one of the remaining 7 variables from model 1013 give a smaller BIC. The LRT for $x_3 = 0$ (benefit type) hypothesis for the Burr model is rejected since $22.4 > 9.7$. Similarly for the lognormal model, the hypothesis $x_4 = 0$ (smoker status) is rejected (LRT $= 12.4 > 9.7$, Table 5.9) on the second step after dropping sex (x_2) from the full model in the first step. Hence model 1013 under the Burr distribution which excludes variables age (x_1), sex (x_2) and smoker status (x_4) and model 1022 under the lognormal model which excludes sex (x_2) are the best models. This result, which is based on a maximum likelihood approach, agrees with the Bayesian variable selection results given in Sections 5.2.1 and 5.2.2.

Table 5.7: Variable selection with backward stepwise method for the Burr model.

Model No	l	BIC	Drop Var
1022	−95160.2	190620.1	x_2
1023	−95162.5	190624.9	x_1
1016	−95162.7	190625.1	x_4
1024	−95158.3	190626.1	$NONE$
1020	−95169.2	190638.2	x_3
1008	−95172.2	190644.2	x_5
960	−95175.0	190649.8	x_7
896	−95273.2	190846.1	x_8
992	−95281.7	190863.2	x_6
768	−95435.3	191063.9	x_9
512	−95429.8	191082.1	x_{10}

Model No	l	BIC	Drop Var (x_2 and .)
1014	−95164.9	190619.9	x_4
1021	−95164.9	190620.0	x_1
1022	−95160.2	190620.1	$NONE$
1018	−95171.1	190632.4	x_3
1006	−95174.1	190638.3	x_5
958	−95177.9	190645.9	x_7
894	−95275.3	190840.7	x_8
990	−95284.3	190858.7	x_6
766	−95437.8	191059.3	x_9
510	−95430.5	191073.8	x_{10}

Model No	l	BIC	Drop Var (x_2, x_4 and .)
1013	−95169.5	190619.4	x_1
1014	−95164.9	190619.9	$NONE$
1010	−95175.5	190631.4	x_3
998	−95178.4	190637.2	x_5
950	−95180.9	190642.2	x_7
886	−95277.8	190836.1	x_8
982	−95288.0	190856.4	x_6
758	−95443.8	191061.6	x_9
502	−95433.4	191070.0	x_{10}

Model No	l	BIC	Drop Var (x_1, x_2, x_4 and .)
1013	−95169.5	190619.4	$NONE$
1009	−95180.7	190632.3	x_3
997	−95181.7	190634.1	x_5
949	−95183.2	190637.3	x_7
981	−95291.0	190852.8	x_6
885	−95295.6	190862.0	x_8
501	−95437.6	191068.6	x_{10}
757	−95466.4	191097.2	x_9

Table 5.8: Variable selection with backward stepwise method for the lognormal model.

Model No	l	BIC	Drop Var
1022	−96060.9	192411.9	x_2
1024	−96060.1	192420.1	$NONE$
1016	−96066.1	192422.3	x_4
1023	−96066.7	192423.6	x_1
1008	−96069.3	192428.7	x_5
1020	−96074.5	192439.2	x_3
960	−96084.2	192458.5	x_7
992	−96174.2	192638.6	x_6
768	−96247.2	192678.1	x_9
896	−96260.0	192810.2	x_8
512	−96373.3	192959.4	x_{10}

Model No	l	BIC	Drop Var (x_2 and .)
1022	−96060.9	192411.9	$NONE$
1014	−96067.1	192414.7	x_4
1021	−96067.9	192416.3	x_1
1006	−96070.0	192420.4	x_5
1018	−96075.3	192431.1	x_3
958	−96085.9	192452.2	x_7
990	−96175.6	192631.6	x_6
766	−96248.2	192670.4	x_9
894	−96261.3	192803.0	x_8
510	−96374.5	192952.0	x_{10}

Table 5.9: LRT values for each step given in Table 5.7 and 5.8.

	Burr				Lognormal	
H_0 :	$x_2 = 0$	$x_4 = 0$	$x_1 = 0$	$x_3 = 0$	$x_2 = 0$	$x_4 = 0$
LRT:	3.8	9.3	9.3	22.4	1.6	12.4

The results obtained so far under different variable selection methodologies give very close answers. Using Bayesian variable selection, marginal likelihoods or ML based methods suggest the same model as the best model, i.e. m_{1013} under the Burr model and m_{1022} under the LN model. DIC values of these models are shown in Table 5.10 and the lower DIC value for the Burr model again suggests that the selected model under the Burr distribution has a better fit than the selected model under a lognormal distribution. Therefore we will continue our analysis using model 1013 under the Burr distribution. Previously, referring to the Bayes factor, we mentioned that under the

Burr distribution we have 4 models ($m_{1013}, m_{1014}, m_{1021}$ and m_{1022}) that are barely different from each other (see Tables 5.3 and 5.6). We select model 1013 instead of the other 3 models because of parsimony. These models are not statistically different from each other and model 1013 is a more parsimonious model compared to other models as it includes less covariate(s) compared to them.

Table 5.10: DIC values of the selected models

	$\bar{D} = -2\widehat{\log L(\boldsymbol{\theta})}$	$\hat{D} = -2\log L(\hat{\boldsymbol{\theta}})$	p_D	**DIC**
Burr (m_{1013})	190367.6	190339.2	28.4	190396.0
LN (m_{1022})	192151.6	192121.8	29.8	192181.4

The posterior and ML estimates of the parameters of the selected model (m_{1013}) under the Burr model are given in Table 5.11.

The earlier comments on the effect of coefficients on the delay distribution are still valid when using the selected model.

Table 5.11: Estimates of the parameters under the selected Burr model (m_{1013}).

	MCMC					**MLE**	
Parameter	Mean	SD	2.5%	50%	97.5%	Mean	SD
β_0	5.2980	0.0268	5.2470	5.2970	5.3490	5.2960	0.0280
β_3	-0.0288	0.0061	-0.0409	-0.0288	-0.0167	-0.0288	0.0061
β_5	0.0312	0.0063	0.0187	0.0313	0.0436	0.0311	0.0063
β_6	0.1140	0.0071	0.1004	0.1139	0.1282	0.1143	0.0073
β_7	-0.0332	0.0065	-0.0459	-0.0331	-0.0208	-0.0331	0.0064
β_8	-0.1202	0.0074	-0.1346	-0.1202	-0.1057	-0.1201	0.0076
$\beta_{9,1}$	0.2406	0.0234	0.1945	0.2404	0.2864	0.2396	0.0239
$\beta_{9,2}$	0.1358	0.0210	0.0944	0.1354	0.1769	0.1345	0.0218
$\beta_{9,3}$	-0.2149	0.0622	-0.3372	-0.2172	-0.0895	-0.2079	0.0611
$\beta_{9,4}$	0.1322	0.0484	0.0408	0.1311	0.2288	0.1342	0.0496
$\beta_{9,5}$	-0.1358	0.0359	-0.2054	-0.1363	-0.0640	-0.1398	0.0368
$\beta_{9,6}$	-0.5528	0.0812	-0.7173	-0.5529	-0.3929	-0.5377	0.0835
$\beta_{9,7}$	-0.2925	0.1270	-0.5539	-0.2841	-0.0520	-0.3003	0.1224
$\beta_{9,8}$	0.0769	0.0213	0.0364	0.0765	0.1192	0.0762	0.0217
$\beta_{9,9}$	-0.1974	0.0262	-0.2485	-0.1975	-0.1452	-0.1989	0.0267
$\beta_{9,10}$	0.2371	0.0318	0.1726	0.2373	0.2976	0.2359	0.0327
$\beta_{9,11}$	-0.0927	0.0193	-0.1300	-0.0930	-0.0541	-0.0942	0.0194
$\beta_{9,12}$	0.1938	0.0253	0.1436	0.1938	0.2451	0.1928	0.0259
$\beta_{9,13}$	0.4697	0.0778	0.3161	0.4733	0.6219	0.4656	0.0778
$\beta_{10,1}$	-0.1523	0.0416	-0.2340	-0.1523	-0.0706	-0.1528	0.0410
$\beta_{10,2}$	-0.0814	0.0208	-0.1211	-0.0817	-0.0402	-0.0808	0.0202
$\beta_{10,3}$	-0.4840	0.0286	-0.5395	-0.4850	-0.4262	-0.4837	0.0281
$\beta_{10,4}$	0.0002	0.0243	-0.0464	0.0002	0.0487	0.0005	0.0240
$\beta_{10,5}$	0.1016	0.0783	-0.0521	0.1014	0.2548	0.0942	0.0803
$\beta_{10,6}$	0.2451	0.1193	0.0190	0.2435	0.4862	0.2565	0.1221
$\beta_{10,7}$	0.1237	0.0336	0.0572	0.1232	0.1898	0.1235	0.0339
$\beta_{10,8}$	0.0167	0.0287	-0.0414	0.0173	0.0718	0.0165	0.0286
$\beta_{10,9}$	0.2347	0.0300	0.1764	0.2342	0.2932	0.2355	0.0290
$\beta_{10,10}$	-0.0044	0.0570	-0.1157	-0.0040	0.1079	-0.0096	0.0603
α	0.6193	0.0150	0.5910	0.6189	0.6485	0.6178	0.0168
τ	2.6280	0.0322	2.5680	2.6280	2.6880	2.6349	0.0382

5.3 Selection of claim-specific covariates with growth factor

In this section claim-specific covariates are selected with Gibbs based methods, using empirical priors and Zellner's g-prior as well as ML based methods under the Burr distribution. We showed that the Burr distribution has a better fit than the LN model in all of the cases we discussed so far. Therefore, here we do not give the full analysis for the LN model but provide some comparisons.

Burr model

Gibbs-based methods:

Since we allow for growth in this section, we introduce $\boldsymbol{\gamma}$ in the model given in (4.2) with the same linear predictor given in Section 5.2.1. That is

$$\eta_i^* = \beta_0 + \sum_{j=1}^{8} \gamma_j \beta_j z_{ij} + \gamma_9 \beta_{9,Office_k} + \gamma_{10} \beta_{10,Cause_l}.$$

We also use the same priors we used in Section 5.2.1 for empirical and Zellner's g-prior. Posterior variable inclusion probabilities $p(\gamma|\boldsymbol{D})$ and posterior model probabilities, $p(m|\boldsymbol{D})$, of 10 models having the highest probabilities with these two prior distribution sets are given in Table 5.12 under the Burr model with growth factor. If we decide to eliminate variables which have an inclusion probability less than 0.5, then age, sex, smoker status and settlement year (x_1, x_2, x_4 and x_6) should be dropped from the model with these two prior sets. This result is consistent with posterior model probabilities, as model 981, which excludes these variables, has the highest probability (see Tables 5.12(b) and 5.12(c)). The second highest model includes age (x_1) and this model (m_{982}) is about 2.5 times less likely than the first model. The top 10 models with both prior sets are exactly the same with changing order for the models which have less than 5% posterior probabilities.

When model selection is carried out using a maximum likelihood based method, the result agrees with the Bayesian variable selection results as model 981 is also found

to be the best model under this approach.

The difference between model 981 and model 1013 is the settlement year. When we did not take the growth factor into account the settlement year was found to be important. However, as we explained in Chapter 4, the effect of settlement year disappears when we allow for the growth between successive years. Since the effect of the year on the delay is partly explained by the growth rate, this result is reasonable.

Table 5.12: Parameter inclusion probabilities and model probabilities under the Burr model with empirical and Zellner's g-priors when business growth is taken into account.

(a) Inclusion probabilities for empirical priors and Zellner's g-prior.

	Empirical Priors		Zellner's g-Prior	
Parameter	$p(\gamma\|\boldsymbol{D})$	SD	$p(\gamma\|\boldsymbol{D})$	SD
γ_1	0.2879	0.4528	0.2987	0.4577
γ_2	0.0841	0.2775	0.0772	0.267
γ_3	0.8538	0.3533	0.8114	0.3912
γ_4	0.256	0.4364	0.2378	0.4258
γ_5	1	0	1	0
γ_6	0.0857	0.28	0.0883	0.2837
γ_7	1	0	0.997	0.0544
γ_8	1	0	1	0
γ_9	1	0	1	0
γ_{10}	1	0	1	0

(b) Posterior model probabilities - empirical priors

Rank	Model No	Model	$p(m\|\boldsymbol{D})$	$PO(m_{981}/mk)$
1	m_{981}	$x_3 + x_5 + x_7 + x_8 + x_9 + x_{10}$	0.3806	1.00
2	m_{982}	$x_1 + x_3 + x_5 + x_7 + x_8 + x_9 + x_{10}$	0.1429	2.66
3	m_{989}	$x_3 + x_4 + x_5 + x_7 + x_8 + x_9 + x_{10}$	0.1388	2.74
4	m_{977}	$x_5 + x_7 + x_8 + x_9 + x_{10}$	0.0624	6.10
5	m_{990}	$x_1 + x_3 + x_4 + x_5 + x_7 + x_8 + x_9 + x_{10}$	0.0613	6.21
6	m_{983}	$x_2 + x_3 + x_5 + x_7 + x_8 + x_9 + x_{10}$	0.0404	9.41
7	m_{1013}	$x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0314	12.13
8	m_{978}	$x_1 + x_5 + x_7 + x_8 + x_9 + x_{10}$	0.0309	12.31
9	m_{1014}	$x_1 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0141	26.94
10	m_{985}	$x_4 + x_5 + x_7 + x_8 + x_9 + x_{10}$	0.0120	31.64

(c) Posterior model probabilities - Zellner's g-Prior

Rank	Model No	Model	$p(m\|\boldsymbol{D})$	$PO(m_{981}/mk)$
1	m_{981}	$x_3 + x_5 + x_7 + x_8 + x_9 + x_{10}$	0.3644	1.00
2	m_{982}	$x_1 + x_3 + x_5 + x_7 + x_8 + x_9 + x_{10}$	0.1496	2.44
3	m_{989}	$x_3 + x_4 + x_5 + x_7 + x_8 + x_9 + x_{10}$	0.1170	3.11
4	m_{977}	$x_5 + x_7 + x_8 + x_9 + x_{10}$	0.0870	4.19
5	m_{990}	$x_1 + x_3 + x_4 + x_5 + x_7 + x_8 + x_9 + x_{10}$	0.0523	6.97
6	m_{978}	$x_1 + x_5 + x_7 + x_8 + x_9 + x_{10}$	0.0355	10.26
7	m_{983}	$x_2 + x_3 + x_5 + x_7 + x_8 + x_9 + x_{10}$	0.0349	10.43
8	m_{1013}	$x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0304	11.99
9	m_{985}	$x_4 + x_5 + x_7 + x_8 + x_9 + x_{10}$	0.0200	18.19
10	m_{1014}	$x_1 + x_3 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$	0.0175	20.82

When we performed model selection under the LN model, we saw that the best model is not changed. Model 1022 which excludes only sex (x_2) from the model is still the best model under the LN distribution when growth rate is introduced. As mentioned in Chapter 4, the importance of the settlement year under the LN model is not very much affected by the introduction of the growth rates. The reason is that the growth rates are affecting the earlier claims more, i.e. they are higher for earlier years for almost all offices (see Table 4.1). Since these earlier claims have longer delays, the higher weights on them increase the importance of modelling the tail. The relatively shorter/inflexible tail of the LN distribution, on the other hand, is not able to model them adequately.

To compare the Burr model with 6 covariates (model 981) with the lognormal model which excludes sex (model 1022), log-likelihood values (l), degrees of freedoms (df) and corresponding BIC values of these models are given in Table 5.13. The considerably lower BIC of the Burr distribution, once again, suggest that it has a better fit. Therefore we select this Burr model (model 981) as the best model when the business growth is allowed in the model.

Table 5.13: Comparison of the Burr and LN models with growth rate.

	Burr (m_{981})	LN (m_{1022})
l	-95610.1	-97146.7
df	28	31
BIC	191491.0	194593.2

We will use the selected model (m_{981}) under the Burr distribution when we introduce missing values in the analysis and for prediction. The posterior and ML estimates of the parameters of this model are given in Table 5.14.

Table 5.14: Estimates of the parameters under the best Burr model (m_{981}) without missing values with growth rate.

			MCMC			MLE	
Parameter	Mean	SD	2.5%	50%	97.5%	Mean	SD
β_0	5.4330	0.0333	5.3660	5.4340	5.4990	5.4389	0.0305
β_3	-0.0227	0.0064	-0.0354	-0.0227	-0.0100	-0.0228	0.0063
β_5	0.0351	0.0065	0.0222	0.0351	0.0478	0.0352	0.0065
β_7	-0.0306	0.0066	-0.0434	-0.0307	-0.0179	-0.0303	0.0065
β_8	-0.0985	0.0075	-0.1133	-0.0984	-0.0843	-0.0981	0.0075
$\beta_{9,1}$	0.3238	0.0247	0.2751	0.3238	0.3714	0.3202	0.0239
$\beta_{9,2}$	0.2665	0.0237	0.2209	0.2664	0.3129	0.2630	0.0225
$\beta_{9,3}$	-0.2103	0.0640	-0.3399	-0.2076	-0.0872	-0.2046	0.0635
$\beta_{9,4}$	-0.2726	0.0494	-0.3677	-0.2725	-0.1788	-0.2759	0.0510
$\beta_{9,5}$	-0.0478	0.0373	-0.1223	-0.0477	0.0233	-0.0499	0.0380
$\beta_{9,6}$	-0.1558	0.0891	-0.3110	-0.1600	0.0321	-0.1504	0.0923
$\beta_{9,7}$	-0.1131	0.1403	-0.3911	-0.1181	0.1465	-0.0952	0.1255
$\beta_{9,8}$	0.1556	0.0230	0.1090	0.1560	0.1995	0.1513	0.0224
$\beta_{9,9}$	-0.3019	0.0281	-0.3550	-0.3026	-0.2493	-0.3068	0.0273
$\beta_{9,10}$	0.2463	0.0347	0.1783	0.2462	0.3150	0.2427	0.0336
$\beta_{9,11}$	-0.1137	0.0212	-0.1556	-0.1132	-0.0745	-0.1186	0.0200
$\beta_{9,12}$	-0.1759	0.0278	-0.2305	-0.1758	-0.1228	-0.1816	0.0262
$\beta_{9,13}$	0.3990	0.0744	0.2637	0.3984	0.5495	0.4058	0.0803
$\beta_{10,1}$	-0.1547	0.0416	-0.2361	-0.1546	-0.0727	-0.1541	0.0420
$\beta_{10,2}$	-0.0715	0.0206	-0.1136	-0.0715	-0.0329	-0.0723	0.0205
$\beta_{10,3}$	-0.4673	0.0291	-0.5226	-0.4678	-0.4092	-0.4685	0.0286
$\beta_{10,4}$	0.0019	0.0251	-0.0467	0.0019	0.0500	0.0019	0.0244
$\beta_{10,5}$	0.1145	0.0788	-0.0415	0.1145	0.2666	0.1151	0.0819
$\beta_{10,6}$	0.2548	0.1287	-0.0022	0.2574	0.5049	0.2582	0.1232
$\beta_{10,7}$	0.1238	0.0355	0.0550	0.1240	0.1936	0.1241	0.0347
$\beta_{10,8}$	0.0231	0.0293	-0.0340	0.0229	0.0803	0.0216	0.0292
$\beta_{10,9}$	0.2438	0.0297	0.1859	0.2443	0.3021	0.2435	0.0296
$\beta_{10,10}$	-0.0684	0.0602	-0.1870	-0.0664	0.0505	-0.0695	0.0606
α	0.5858	0.0156	0.5581	0.5866	0.6156	0.5840	0.0158
τ	2.6250	0.0378	2.5600	2.6200	2.7060	2.6283	0.0387

Discussion

In this chapter, the most relevant covariates affecting the delay distribution are chosen first without considering business growth and then taking it into account. Sex is not significant throughout the analyses under the Burr and LN models. On the other hand, leaving age and smoker status out of the model is ambiguous since they have relatively high posterior inclusion probabilities under the Burr model. We expect the effect of year to diminish when office specific growth rates between successive years are introduced in the model. Typically in life-related insurance practice, age and smoking status are important policyholder characteristics and should probably also

be taken into consideration when inception rates are calculated in CII. We note that our model is developed here with the purpose of estimating and predicting delay in claim settlement, also in the presence of non-recorded dates of diagnosis or settlement.

Chapter 6

Modelling CDD III: Including the missing values

6.1 Introduction

The data contain 19127 delay observations which have either the date of diagnosis or the date of settlement recorded, with 15860 of them having both dates. Hence 17% of the claims are excluded from the analysis when records with missing information are not taken into account. However, in Bayesian analysis, the posterior distribution of a parameter vector $\boldsymbol{\theta}$ can be obtained by conditioning only on observed values $\boldsymbol{D}_{obs}$ (Gelman *et al.*, 2000). The aim of this chapter is to estimate the CDD taking account of business growth and also missing values in the data set. This can then provide estimates of non-recorded dates that are important for modelling claim inception rates.

The growth rates are assigned to claims according to their diagnosis years. This means that growth rates of 1752 claims (9.2%) which do not have date of diagnosis recorded, can not be directly calculated. These claims are included in the analyses in two steps. First, we estimate the dates of diagnosis from the model which assumes no growth within offices between consecutive years. Here, we use the selected Burr model derived in Section 5.2. This step is described in Section 6.2. Then, in the second step, we use these dates to assign growth rates to the claims with missing dates of diagnosis

and re-estimate the missing values allowing for the growth in the offices. The model is developed in Section 6.3. In this second step, the selected Burr model concluded in Section 5.3 is employed.

6.2 Including the missing values and assuming no growth within offices

The dependent variable vector, $\boldsymbol{D}$, can be written as $\boldsymbol{D} = (\boldsymbol{D}_{obs}, \boldsymbol{D}_{mis})$ where $\boldsymbol{D}_{obs}$ denotes the observed values and $\boldsymbol{D}_{mis}$ denotes the missing values. So, the joint density of the missing delay values conditional on the observed data can be given by

$$\begin{aligned} P\left(\boldsymbol{D}_{mis}|\boldsymbol{D}_{obs}\right) &= \int P\left(\boldsymbol{D}_{mis}, \boldsymbol{\theta}|\boldsymbol{D}_{obs}\right) d\boldsymbol{\theta} \\ &= \int P\left(\boldsymbol{D}_{mis}|\boldsymbol{\theta}, \boldsymbol{D}_{obs}\right) P\left(\boldsymbol{\theta}|\boldsymbol{D}_{obs}\right) d\boldsymbol{\theta} \\ &= \int P\left(\boldsymbol{D}_{mis}|\boldsymbol{\theta}\right) P\left(\boldsymbol{\theta}|\boldsymbol{D}_{obs}\right) d\boldsymbol{\theta} \\ &= \underset{\boldsymbol{\theta}|\boldsymbol{D}_{obs}}{E}\left[P\left(\boldsymbol{D}_{mis}|\boldsymbol{\theta}\right)\right]. \end{aligned}$$

This suggests that the density of the missing values can be estimated using the simulated $\boldsymbol{\theta}$ values from the MCMC output. For each missing observation, values are drawn from its posterior distribution. On the other hand, when we have missing observations in the independent variables, full conditional distributions should be specified, i.e. we need a distribution for those missing observations. By doing this, missing values will be introduced as parameters to be estimated in WinBUGS (Ntzoufras, 2009).

In addition to dates of diagnosis (DoD) and settlement (DoS), we are also provided with dates of commencement (DoC) of the policy, notification (DoN) and claim admission (DoA) related to each claim. Since these dates are expected to be in chronological order (see Figure 6.1), we use them to obtain natural upper and lower limits when we impute the missing delay values, i.e.

$$D_u \sim Burr(\alpha, \tau, \lambda_u) I(LB_u, UB_u) \tag{6.1}$$

where $u = 15861, ..., 19127$, and LB and UB are lower and upper bounds, respectively.

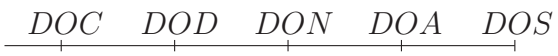


Figure 6.1: Chronological order of the dates relating to a claim.

The lower bound is determined as follows

$d_u \geq DoS - DoN$ when DoD is missing,

$d_u \geq DoA - DoD$ when DoS is missing,

$d_u \geq DoN - DoD$ when DoS and DoA are missing,

$d_u > 0$ when there is only DoD or DoS recorded.

The upper bound of the delay is determined by $DoS - DoC \geq d_u$ for the claims where DoD is not recorded.

Due to missing dates of diagnosis, unobserved values of the covariate giving policy duration until diagnosis must be imputed as well. We need to assign appropriate distributions as follows

$$\begin{aligned} z_{PolDur_v} &\sim N(\mu_{PolDur}, \sigma^2_{PolDur})\text{I(-1.22, 6.17)} \\ \mu_{PolDur} &\sim N(0, 1000) \\ \sigma^2_{PolDur} &\sim IGa(0.001, 0.001) \end{aligned} \tag{6.2}$$

for $v = 17376, ..., 19127$. This means that we use a lower bound of zero days with an upper bound of 7000 days resulting in the $I(-1.22, 6.17)$ restriction for the standardised covariate.

We use the selected Burr model (m_{1013}) in Section 5.2. The model has the same structure given in (3.5), however, here we also introduce missing values in the model, i.e.

$$D_i \sim Burr(\alpha, \tau, \lambda_i), \qquad i = 1, \ldots, 15860$$

$$D_u \sim Burr(\alpha, \tau, \lambda_u) I(LB_u, UB_u), \qquad u = 15861, \ldots, 19127$$

and for the model parameters, we use the prior distributions given in (3.10) but, here we exclude age (x_1), sex (x_2) and smoker status (x_4) from the full model. For the unobserved values of policy duration, we use the distribution given in (6.2). The estimated means of the standardised coefficients, their standard deviations and credible intervals are shown in Table 6.1, together with estimates of the parameters of the policy duration distribution. Note that the definitions of the covariates corresponding to β coefficients can be seen in Table 3.3.

We can compare the posterior estimates with those given in Table 5.11 and changes can also be seen in Figures 6.2 and 6.3. These figures show the posterior densities of the model parameters under the Burr model with and without missing values. Including missing observations in the analysis leads to significant changes in the posterior densities of some of the model parameters which means that for these covariates the missing data are systematically different from the observed data. For example including missing observations in the analysis decreases the positive effect of settlement year (β_6) on the delay. It might be because data are not provided for some settlement years when the delay is shorter. The negative effect of Office 6 ($\beta_{9,6}$) decreases when the missing information is taken into account. The reason might be the office does not provide data when the delay is longer. For cancer claims, on the other hand, the data might not be provided when the delay is shorter, so that including the missing data increases the negative effect of cancer ($\beta_{10,2}$) on the delay. Other changes can be interpreted in a similar way. The sign of the mean estimates of heart attack ($\beta_{10,4}$) and other causes ($\beta_{10,8}$) changes from positive to negative. However in both cases these coefficients are not significant in the models. The sign of TPD ($\beta_{10,10}$) changes from negative to positive when we include the missing observations in the model. Note that this coefficient was negative with a very high standard deviation in the analysis

without missing values. This might mean that for the TPD claims, data are provided when the delays are shorter.

Table 6.1: Posterior estimates of parameters under the selected Burr model with missing values (m_{1013}).

Parameter	Mean	SD	2.5%	50%	97.5%
β_0	5.3690	0.0277	5.3100	5.3690	5.4240
β_3	-0.0314	0.0060	-0.0432	-0.0314	-0.0196
β_5	0.0325	0.0063	0.0203	0.0326	0.0447
β_6	0.0937	0.0072	0.0797	0.0936	0.1076
β_7	-0.0329	0.0064	-0.0455	-0.0330	-0.0204
β_8	-0.1151	0.0074	-0.1294	-0.1152	-0.0999
$\beta_{9,1}$	0.2452	0.0228	0.2008	0.245	0.2903
$\beta_{9,2}$	0.1553	0.021	0.1146	0.1552	0.1953
$\beta_{9,3}$	-0.2221	0.0588	-0.3351	-0.2201	-0.1141
$\beta_{9,4}$	0.1390	0.0488	0.0462	0.1381	0.2348
$\beta_{9,5}$	-0.1812	0.0358	-0.2511	-0.1809	-0.1082
$\beta_{9,6}$	-0.4045	0.0764	-0.5549	-0.4011	-0.2564
$\beta_{9,7}$	-0.3327	0.1292	-0.5829	-0.3263	-0.0812
$\beta_{9,8}$	0.0470	0.0213	0.0052	0.0472	0.0885
$\beta_{9,9}$	-0.2182	0.0246	-0.2675	-0.2177	-0.1705
$\beta_{9,10}$	0.1845	0.0317	0.1236	0.1841	0.2477
$\beta_{9,11}$	-0.1370	0.0188	-0.1735	-0.1375	-0.0996
$\beta_{9,12}$	0.1511	0.0244	0.1022	0.1513	0.1981
$\beta_{9,13}$	0.5735	0.0513	0.4729	0.5746	0.6731
$\beta_{10,1}$	-0.1602	0.0398	-0.2394	-0.1603	-0.0829
$\beta_{10,2}$	-0.1269	0.0202	-0.1676	-0.1266	-0.0868
$\beta_{10,3}$	-0.4576	0.0264	-0.5075	-0.4579	-0.4038
$\beta_{10,4}$	-0.0471	0.0237	-0.0924	-0.0471	0.0001
$\beta_{10,5}$	0.0848	0.0795	-0.0665	0.0845	0.2446
$\beta_{10,6}$	0.2339	0.1163	-0.0002	0.2343	0.4572
$\beta_{10,7}$	0.1324	0.0328	0.0682	0.1329	0.1954
$\beta_{10,8}$	-0.0146	0.0288	-0.0696	-0.0153	0.0431
$\beta_{10,9}$	0.1902	0.0277	0.1359	0.1898	0.246
$\beta_{10,10}$	0.1652	0.0547	0.0538	0.1675	0.2666
α	0.6403	0.0181	0.6030	0.6415	0.6723
τ	2.5790	0.0367	2.5150	2.5750	2.6600
μ_{PolDur}	0.0188	0.0074	0.0043	0.0188	0.0333
$(\sigma^2_{PolDur})^{-1}$	1.0260	0.0108	1.0050	1.0260	1.0480

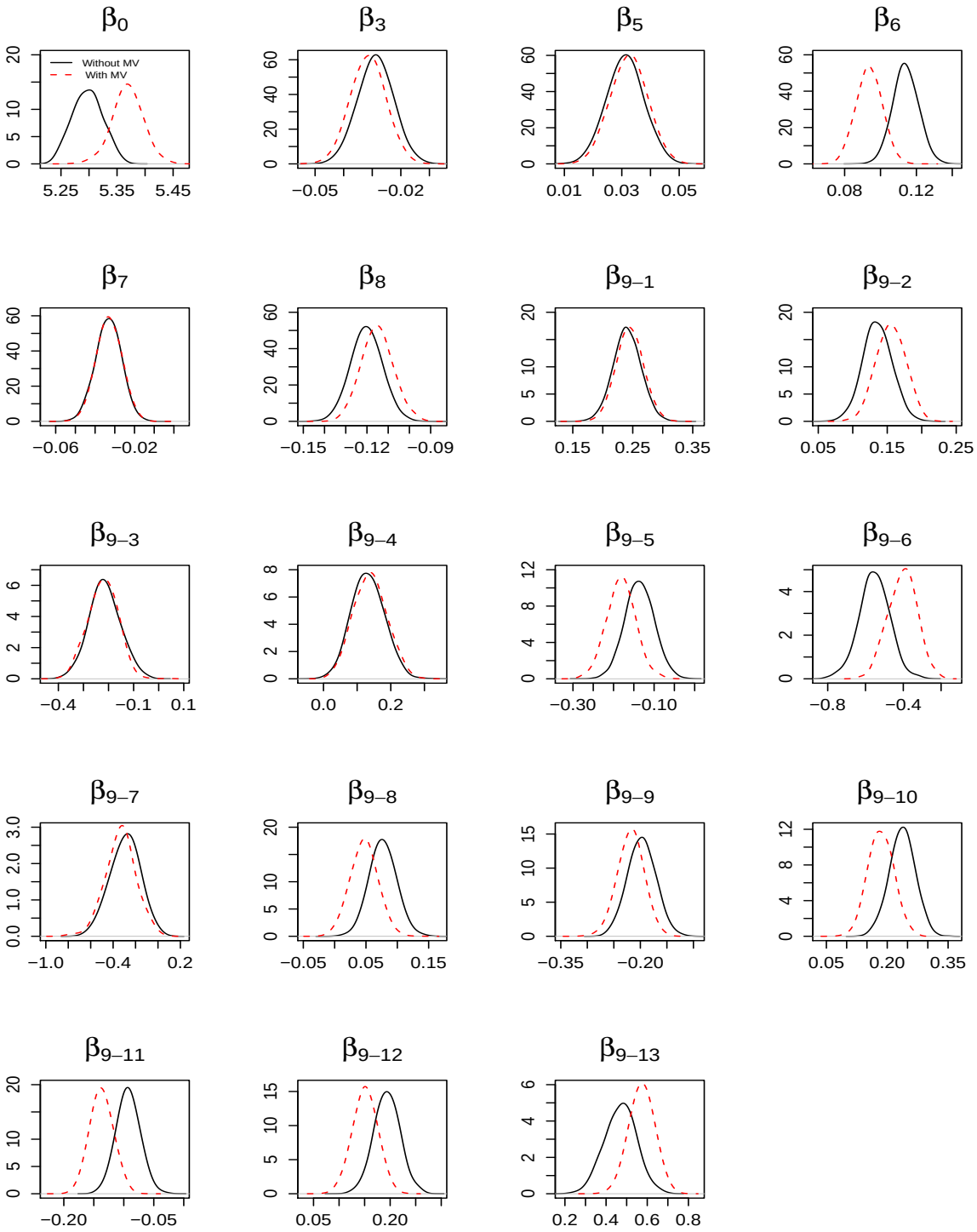


Figure 6.2: Comparison of posterior densities of model parameters ($\beta_1 - \beta_9$) under the selected Burr model (m_{1013}) with (red dashed line) and without (black solid line) missing values.

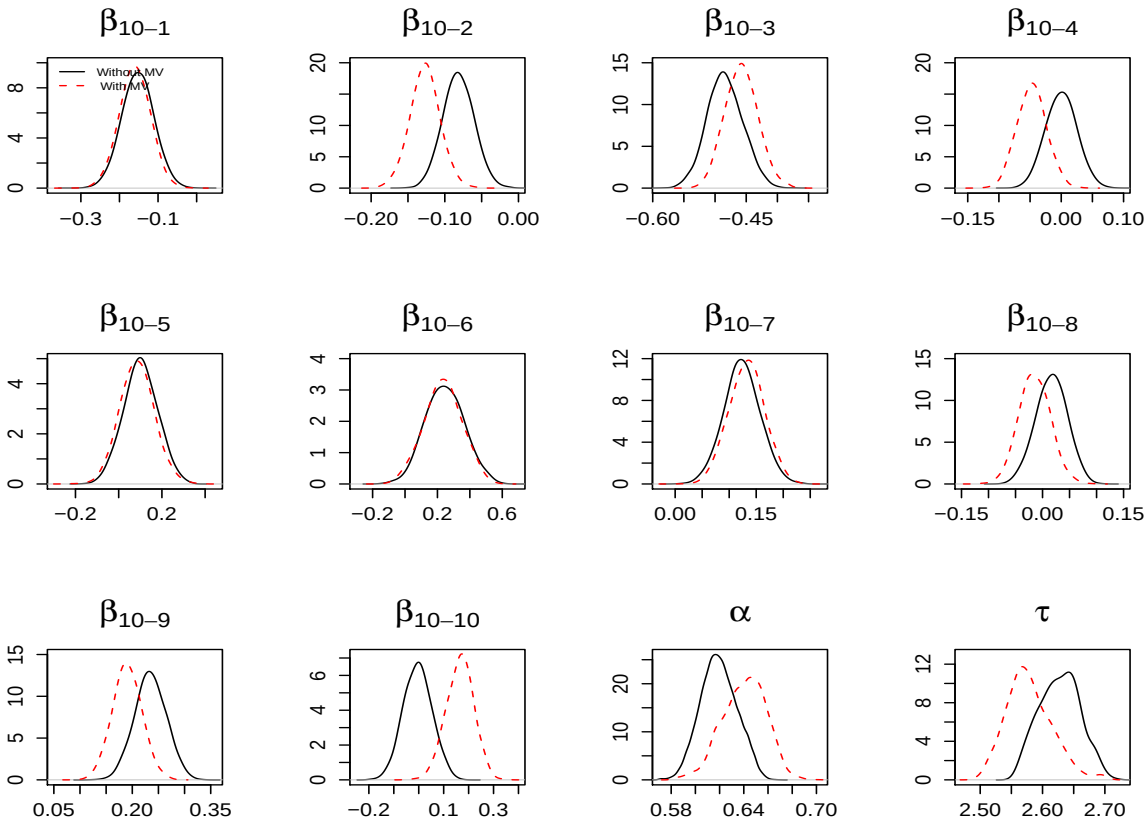


Figure 6.3: Comparison of posterior densities of model parameters $(\beta_{10}, \alpha, \tau)$ under the selected Burr model (m_{1013}) with (red dashed line) and without (black solid line) missing values.

6.3 Considering the missing values and the growth rate

In this section, both the missing values and the growth rates are handled in the delay distribution. The modelling of the delay is performed in four steps. First, the missing dates of diagnosis are filled with the estimated delays from the model in Section 6.2 (step 1). The estimate we use is the date of settlement minus the median of the delay distribution corresponding to the characteristics of the claim. Then, using the year of diagnosis, initial growth rates are assigned to the claims where the *DoD* is missing

(9.2% of the claims) (step 2) and the delay is modelled using the selected model in Section 5.3 (step 3). These 3 steps are repeated once more (starting from the model estimated in step 3) to have the final growth rates and the final CDD (step 4). These steps can be summarised in the following way

(1): Estimate year of diagnosis from the model without growth rate (given in Section 6.2).

(2): Assign growth rates based on (1).

(3): Fit CDD with growth rate from (2) with the selected model presented in Section 5.3.

(4): Repeat (1) - (3) starting from the model obtained in (3). (That is estimate year of diagnosis from (3) and calculate growth rates based on this model. Fit CDD with these growth rates using the model presented in Section 5.3.)

Out of 1752 claims, year of diagnosis (and hence the growth rate) is changed for 52 claims after this second iteration (step 4) either by going one year up or down. Here, model m_{981} (excludes age, sex, smoker status and year of settlement) is used as it is the selected model according to the variable selection in Section 5.3.

The probability model structure used here is the same as in (4.2). Here, we also include the missing observations in the model. That is

$$D_i \sim Burr(\alpha, \tau, \lambda_{w_i}), \qquad i = 1, \ldots, 15860$$

$$D_u \sim Burr(\alpha, \tau, \lambda_{w_u}) I(LB_u, UB_u), \qquad u = 15861, \ldots, 19127$$

and prior distributions assigned to the model parameters are the same as in (4.3) (note that we exclude age (x_1), sex (x_2), smoker status (x_4) and settlement year (x_6) from the full model) and for the unobserved values of the policy duration covariate, we use the distributions given in (6.2). Posterior estimates of the model parameters are given in Table 6.2 for the Burr model.

Table 6.2: Coefficients of the Burr model (m_{981}) with missing values and with growth rate.

Parameter	Mean	SD	2.5%	50%	97.5%
β_0	5.4690	0.0250	5.4170	5.4690	5.5170
β_3	-0.0234	0.0062	-0.0353	-0.0234	-0.0115
β_5	0.0339	0.0063	0.0218	0.0340	0.0463
β_7	-0.0320	0.0065	-0.0445	-0.0321	-0.0193
β_8	-0.0977	0.0072	-0.1119	-0.0978	-0.0836
$\beta_{9,1}$	0.3031	0.0215	0.2612	0.3030	0.3454
$\beta_{9,2}$	0.2154	0.0196	0.1770	0.2153	0.2543
$\beta_{9,3}$	-0.2052	0.0611	-0.3304	-0.2038	-0.0920
$\beta_{9,4}$	-0.2487	0.0496	-0.3454	-0.2483	-0.1499
$\beta_{9,5}$	-0.0899	0.0368	-0.1638	-0.0894	-0.0182
$\beta_{9,6}$	-0.0504	0.0847	-0.2140	-0.0512	0.1142
$\beta_{9,7}$	-0.1293	0.1182	-0.3741	-0.1240	0.0914
$\beta_{9,8}$	0.1055	0.0205	0.0659	0.1055	0.1462
$\beta_{9,9}$	-0.3154	0.0253	-0.3661	-0.3152	-0.2668
$\beta_{9,10}$	0.2009	0.0329	0.1364	0.2012	0.2662
$\beta_{9,11}$	-0.1576	0.0171	-0.1920	-0.1581	-0.1240
$\beta_{9,12}$	-0.2089	0.0231	-0.2547	-0.2083	-0.1639
$\beta_{9,13}$	0.5805	0.0470	0.4893	0.5799	0.6720
$\beta_{10,1}$	-0.1452	0.0402	-0.2238	-0.1452	-0.0664
$\beta_{10,2}$	-0.1009	0.0183	-0.1364	-0.1010	-0.0659
$\beta_{10,3}$	-0.5419	0.0264	-0.5942	-0.5417	-0.4896
$\beta_{10,4}$	-0.0288	0.0228	-0.0727	-0.0290	0.0162
$\beta_{10,5}$	0.1287	0.0787	-0.0324	0.1289	0.2872
$\beta_{10,6}$	0.1939	0.1163	-0.0398	0.1945	0.4142
$\beta_{10,7}$	0.1523	0.0331	0.0868	0.1524	0.2162
$\beta_{10,8}$	0.0063	0.0277	-0.0472	0.0061	0.0623
$\beta_{10,9}$	0.2149	0.0271	0.1620	0.2145	0.2680
$\beta_{10,10}$	0.1207	0.0564	0.0097	0.1211	0.2313
α	0.6179	0.0153	0.5930	0.6161	0.6520
τ	2.5700	0.0344	2.4940	2.5750	2.6250
μ_{PolDur}	0.0190	0.0074	0.0047	0.0190	0.0333
$(\sigma^2_{PolDur})^{-1}$	1.0260	0.0108	1.0050	1.0260	1.0470

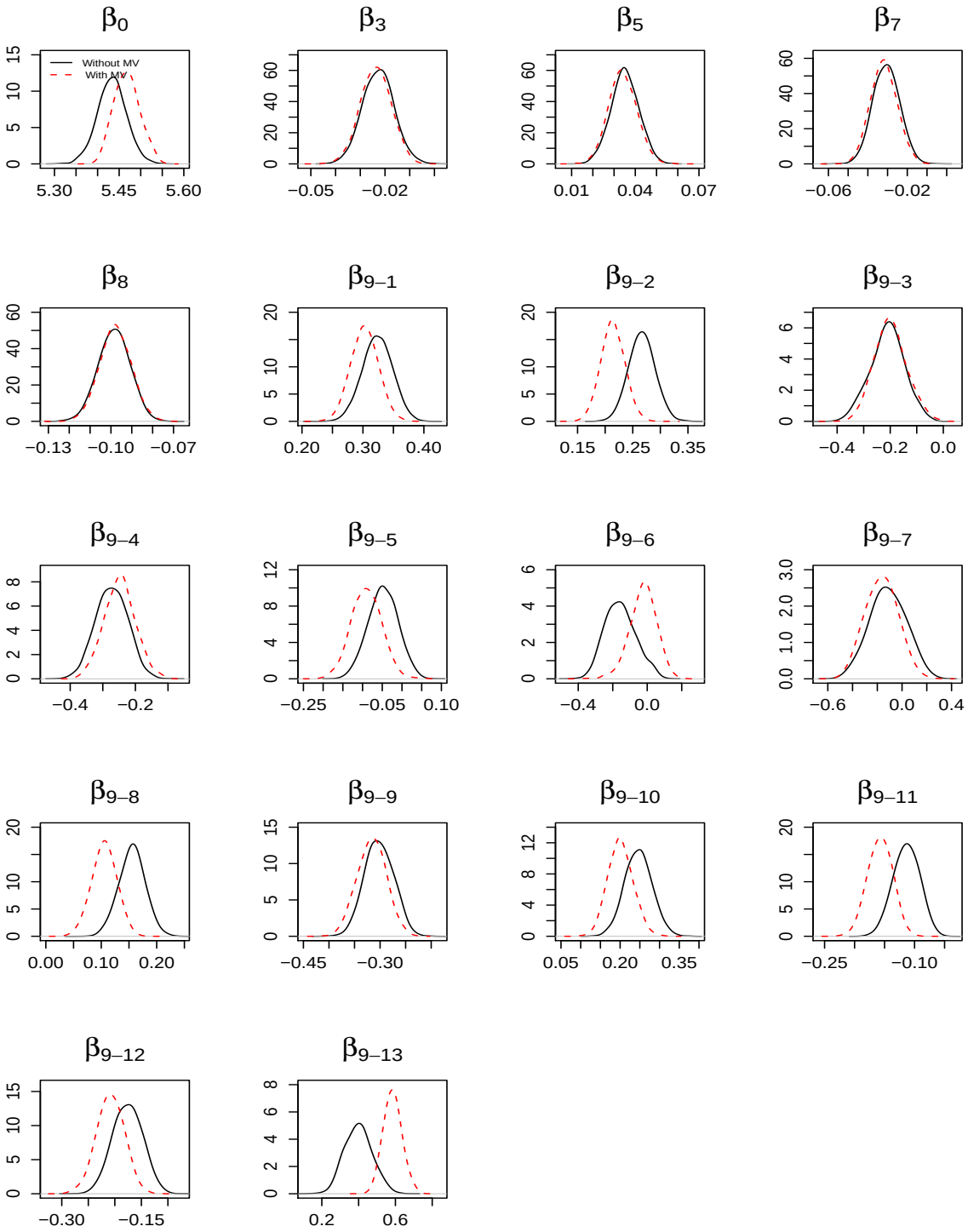


Figure 6.4: Comparison of posterior densities of model parameters ($\beta_1 - \beta_9$) under the selected Burr model (m_{981}) including the growth rate, with (red dashed line) and without (black solid line) missing values.

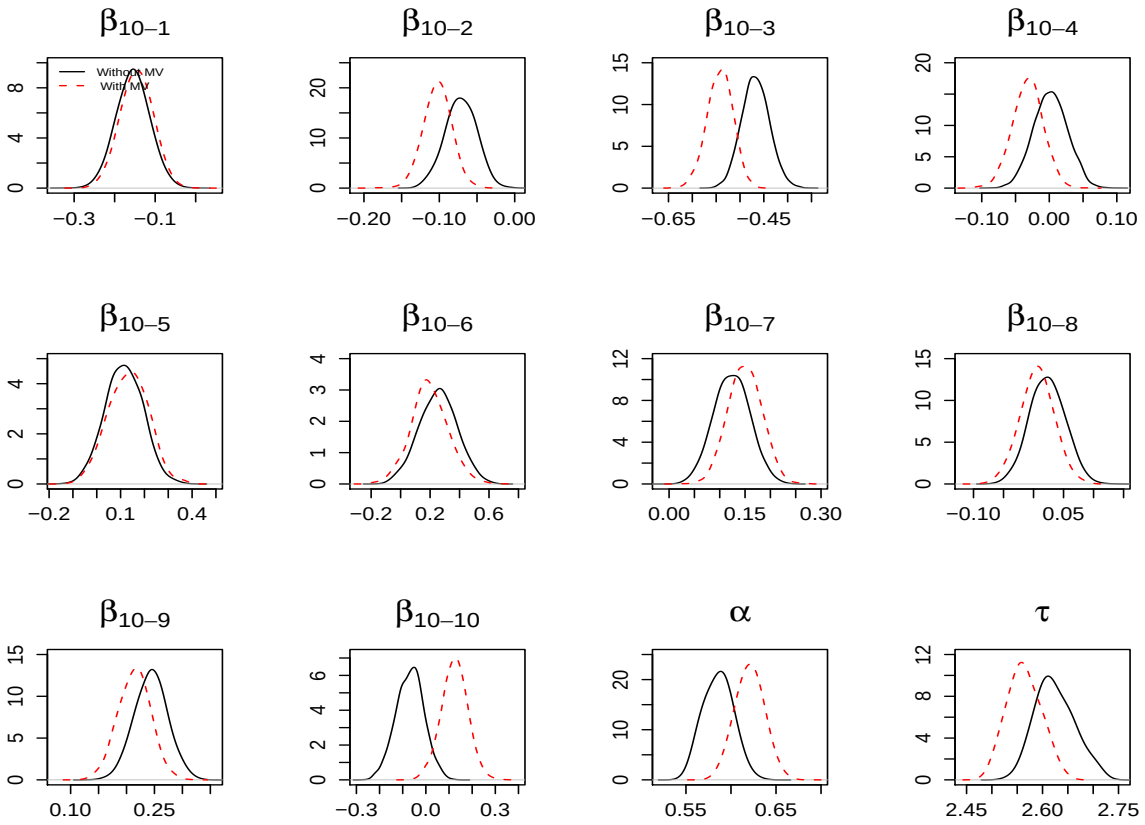


Figure 6.5: Comparison of posterior densities of model parameters $(\beta_{10}, \alpha, \tau)$ under the selected Burr model (m_{981}) including the growth rate, with (red dashed line) and without (black solid line) missing values.

Including missing observations and growth rates in the analysis has an effect on some office (β_9) and cause (β_{10}) levels. Figures 6.4 and 6.5 show the differences in the posterior densities of model parameters with and without missing values. According to Figure 6.4, posterior densities of the effect of Office 13 $(\beta_{9,13})$ are different from each other. One of the reasons for this could be the percentage of missing observations for this office, 86.4%, which is the highest among all the offices. Next, Office 2 and 3 come with 28.6% and 25.6% respectively. The posterior densities of Office 3 $(\beta_{9,3})$ are almost exactly the same when the missing values are included or excluded suggesting the missing data provided from this office are 'random', i.e. the missing data are not systematically different from the observed cases. For the causes, the highest missing claims percentage is for death $(\beta_{10,3})$ with 71.4% of the claims not recorded. Con-

sidering the shapes of the posterior densities, there might be a systematic difference between the observed data and the missing data for death. For TPD ($\beta_{10,10}$) the posterior densities are found to be different indicating the missing data might depend on a factor. The data might not be provided when the delay is long for TPD claims.

Prediction

After the variable selection performed in Section 5.3, we decided to exclude age, sex, smoker status and year of settlement from the model. Considering this result, Table 6.3 shows 11 illustrative scenarios and Table 6.4 shows details of the estimated claim delay distribution for each of these scenarios under the Bayesian approach.

Table 6.3: Scenarios for prediction of the CDD under the selected Burr model (m_{981}).

Scenario	1	2	3	4	5	6
Benefit Type	FA	**SA**	FA	FA	FA	FA
Joint/Single Life	J	J	**S**	J	J	J
Benefit Amount	50000	50000	50000	**10000**	**250000**	50000
Policy Duration	1460	1460	1460	1460	1460	**365**
Office Code	11	11	11	11	11	11
Cause of Claim	Cancer	Cancer	Cancer	Cancer	Cancer	Cancer

Scenario	7	8	9	10	11
Benefit Type	FA	FA	FA	FA	FA
Joint/Single Life	J	J	J	J	J
Benefit Amount	50000	50000	50000	50000	50000
Policy Duration	**3650**	1460	1460	1460	1460
Office Code	11	**6**	**10**	11	11
Cause of Claim	Cancer	Cancer	Cancer	**Death**	**TPD**

The mean delay for a typical risk profile (Scenario 1) is estimated as 180 days. Including missing observations in the analysis, the prediction becomes 174 days. Changing the office in the typical scenario to Office 10 (Scenario 9) gives the highest posterior mean delay among these 11 scenarios with 259 days (249 days when missing values are included) while changing the claim cause to death (Scenario 10) gives the shortest one with 121 days (112 days with missing values). Mean delays of the scenarios with their 95% credible intervals for the cases where claims with missing information are included or excluded from the model are given in Figure 6.6. Except for Scenarios 8

and 11, means of the delays considered here are shorter when missing values are taken into account. Credible intervals of these scenarios are wider compared to the others.

Table 6.4: Posterior estimates of the mean of the delay distribution under different scenarios given in Table 6.3 using the selected Burr model (m_{981}) with growth rates.

	Excluding Missing Data					**Including Missing Data**				
	Mean	SD	2.5%	50%	97.5%	Mean	SD	2.5%	50%	97.5%
Mean.Scn1	180.3	4.5	171.4	180.2	189.3	173.9	4.0	166.5	173.8	181.7
Mean.Scn2	168.1	5.3	157.9	168.0	178.9	161.7	4.8	152.7	161.7	171.3
Mean.Scn3	193.4	5.0	183.9	193.4	203.4	186.1	4.3	177.8	186.0	194.9
Mean.Scn4	184.2	4.7	175.1	184.2	193.4	177.8	4.1	170.2	177.8	186.0
Mean.Scn5	162.0	5.7	151.0	161.9	173.3	155.5	5.1	146.0	155.3	166.1
Mean.Scn6	202.1	5.1	192.0	202.0	212.2	194.7	4.4	186.5	194.7	203.5
Mean.Scn7	143.6	4.7	134.7	143.5	152.8	138.7	4.2	130.7	138.6	147.1
Mean.Scn8	173.6	17.3	145.8	171.5	212.6	194.3	17.8	161.6	193.5	230.9
Mean.Scn9	258.5	10.6	237.4	258.4	279.1	249.0	10.0	230.0	248.7	269.7
Mean.Scn10	121.4	3.8	114.2	121.3	128.9	111.9	3.2	105.7	111.8	118.5
Mean.Scn11	181.1	11.2	160.2	180.9	204.0	217.4	12.7	193.4	217.3	242.6

Discussion

Our principal application for this CDD model is to estimate the date of diagnosis (or date of settlement) for a claim record where this is missing. If both dates are missing, the claim record is omitted from the analysis. Hence, all our claim records will have either actual or estimated dates of diagnosis and settlement. By estimating the date of diagnosis, those claims can be used in calculation of inception rates. The effect of using a point estimate (i.e. median) for the missing delays on the inception rates might be argued. In Section 7.5, it is shown that the inception rates are not very sensitive to how the missing dates are estimated.

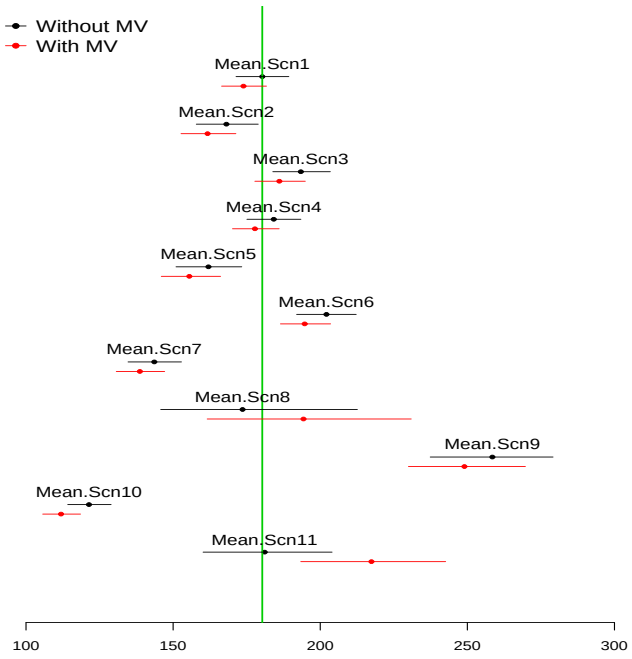


Figure 6.6: Comparison of posterior estimates of the mean delay under different scenarios using the selected Burr model (m_{981}) including (red solid line) and excluding (black solid line) the missing information. Bars show 95% credible intervals and bullets show posterior means.

Chapter 7

Diagnosis inception rates I: All-cause rates

7.1 Introduction

The objective of this chapter is to estimate the intensity of a diagnosis, denoted λ, of a sickness at age x last birthday which will lead to a claim from any cause. This intensity could be a function of some or all of the characteristics listed in Table 7.1.

For the analyses we restrict the age to $16 - 80$. By doing so, we exclude 8 cases out of 587 177 different combinations of profiles where there is exposure at risk. Benefit amount is modelled as a factor and the boundaries for the categories in Table 7.1 are determined appoximately by the quartiles of this covariate. We use the corresponding in force data, for which we have start of year and end of year information for 1999 to 2005 supplied by the CMI. This gave 24 132 215 policy-years of exposure. We omitted approximately 5 million policy years of exposure from these data for various reasons, in particular because some offices do not give information about both dates of diagnosis and settlement for their claims and some policies have undefined smoker status.

In Section 7.2 we explain how we calculated the exposure and present the model we used to estimate the diagnosis inception rates. In Section 7.3 the diagnosis inception rates are estimated under the best model after variable selection. The CMI gives the

Table 7.1: Definitions of the covariates used in the modelling of the intensity rates.

	Covariate	Number of Levels	Additional Information
θ_1	Sex	2 (F & M)	F is the base category
θ_2	Benefit type	2 (FA & SA)	FA is the base category
θ_3	Smoker status	2 (N & S)	N is the base category
θ_4	Policy type	2 (Joint/Single life)	J is the base category
θ_5	Year	Numerical $(1999, \ldots, 2005)$	Calendar year of exposure Calendar year of diagnosis
θ_6	Benefit amount	4	1, Benefit amount<25000 2, 25000<Benefit amount<50000 3, 50000<Benefit amount<75000 4, Benefit amount>75000
θ_7	Policy duration	6	duration between the beginning of the year and commencement of the policy Duration 0, Policy Duration $<$ 1 year Duration 1, 1 year $<$ Policy Duration $\leq$ 2 years Duration 2, 2 years $<$ Policy Duration $\leq$ 3 years Duration 3, 3 years $<$ Policy Duration $\leq$ 4 years Duration 4, 4 years $<$ Policy Duration $\leq$ 5 years Duration 5+, Policy Duration $>$ 5 years
θ_8	Office	13	

diagnosis inception rates for full accelerated policies only and the rates are separated for sex, smoker status and policy duration. So, to compare our rates with CMI rates presented in WP 43 (2010) we needed a model including these covariates. For this purpose, we estimated inception rates with two other models. The details of the CMI variables and difference between these two models are explained in Section 7.4 and for different combinations of characteristics comparisons between the rates are presented. Finally in Section 7.5, we performed a sensitivity analysis to see whether there is a significant difference between the inception rates estimated by using the median of the CDD obtained in Chapter 6 of the estimated delays rather than using other percentiles of this distribution as a point estimate.

7.2 Modelling techniques

Exposure to risk calculation

The in force data we have are in census form, obtained at 1 January and 31 December of each calendar year between 1999 and 2005. The total number of policies inforce may be different for the end (31 December) of one year and beginning (1 January) of the

next. Changes in participation of contributing offices across years are the main reason for this. Also, for some years during their contribution periods, some of the offices have different figures for the end of year and the start of the next year if the office changes its portfolio of policies during its contribution period. However the effect of the latter is relatively small. We also note that there are no gaps in submission, i.e. there are no offices which stop contributing and then start again after a while.

The beginning and end of year in force data are counted for each calendar year for different risk profiles. At the beginning of a year we count the lives at age x last birthday at 1 January ($E^{(0)}(x, \boldsymbol{\theta})$) and at the end of a year we count the lives at age x last birthday at 31 December ($E^{(1)}(x, \boldsymbol{\theta})$) for a given $\boldsymbol{\theta}$ denoting a risk profile involving specific characteristics given in Table 7.1. We assume linear change of exposure between census dates and use a repeated Simpson's Rule for approximation as explained later in this section.

Estimation of the inception rates

Let

t_{θ_8} denote the final calendar year in which the considered office (θ_8) contributes data, so that t_{θ_8} is one of $1999, \ldots, 2005$

$N(x, \boldsymbol{\theta})$ denote the observed number of claims for all-causes at age x and risk profile $\boldsymbol{\theta}$ which are diagnosed in year θ_5 and settled before the end of the last contribution year of the considered office, t_{θ_8}

$E(x, u; \boldsymbol{\theta})$ denote the number of policies with age x last birthday and risk profile $\boldsymbol{\theta}$, exposed at time u in the considered calendar year (θ_5)

$F(t_{\theta_8} - \theta_5 + 1 - u; x, \boldsymbol{\theta})$ denote the probability of settling a claim in time $(t_{\theta_8} - \theta_5 + 1 - u)$ at age x, given $\boldsymbol{\theta}$ by the end of the last contribution year of the considered office, t_{θ_8}, starting from time u after the start of year θ_5, where $\theta_5 = 1999, \ldots, 2004, 2005$ and $t_{\theta_8} \in (\theta_5, \theta_5 + 1, \ldots, 2005)$

$\lambda(x, \boldsymbol{\theta})$ denote the intensity rate (hazard) for a claim diagnosis (all-causes) in the respective year θ_5 at age x last birthday given risk profile $\boldsymbol{\theta}$.

Note that the risk profile, $\boldsymbol{\theta}$, includes office. Any claim settled before the end of the last contribution year t_{θ_8} which has a year of diagnosis θ_5, where the associated office did not contribute data, should be removed from $N(x, \boldsymbol{\theta})$. There are 900 such cases which correspond to 4.7% of the claims data.

We assume the observed number of claim counts has a Poisson distribution given as

$$N(x, \boldsymbol{\theta}) \sim Poisson(\lambda(x, \boldsymbol{\theta}) E^*(x, u; \boldsymbol{\theta})) \tag{7.1}$$

where

$$E^*(x, u; \boldsymbol{\theta}) = \int_{u=0}^{1} E(x, u; \boldsymbol{\theta})\, F(t_{\theta_8} - \theta_5 + 1 - u; x, \boldsymbol{\theta})\, du. \tag{7.2}$$

Here $E^*(x, u; \boldsymbol{\theta})$ can be regarded as an 'adjusted exposure' where the cdf of the appropriate CDD under the Burr model which takes business growth into account, $F(t_{\theta_8} - \theta_5 + 1 - u; x, \boldsymbol{\theta})$, is used as an adjustment factor. The necessity of such an adjustment is explained in Section 1.2. The fact that critical illness claims can be subject to long delays between diagnosis and settlement, as discussed in earlier chapters, makes it essential to use. Therefore, $F(t_{\theta_8} - \theta_5 + 1 - u; x, \boldsymbol{\theta})$ is used as an adjustment factor. It eliminates the distortion caused by the addition of claims diagnosed in contributing years of the office but settled in later years or yet to be settled.

For a given year of diagnosis, θ_5, and a very small interval du ($du \to 0$), the expected number of diagnoses between u and $u + du$ is $\lambda(x, \boldsymbol{\theta})\, E(x, u; \boldsymbol{\theta})\, du$, and the expected number of diagnoses between u and $u + du$ which will be settled before the end of t_{θ_8} is

$$\lambda(x, \boldsymbol{\theta})\, E(x, u; \boldsymbol{\theta})\, F(t_{\theta_8} - \theta_5 + 1 - u; x, \boldsymbol{\theta})\, du.$$

Then the expected number of claims diagnosed in year θ_5 and settled before the end of year t_{θ_8} is

$$\int_{u=0}^{1} \lambda(x, \boldsymbol{\theta})\, E(x, u; \boldsymbol{\theta})\, F(t_{\theta_8} - \theta_5 + 1 - u; x, \boldsymbol{\theta})\, du \tag{7.3}$$

(see Figure 7.1).

Under the Poisson distribution in (7.1) we obtain a raw/crude estimator for the intensity ($\hat{\lambda}^{raw}(x, \boldsymbol{\theta})$), given by the MLE of λ as

$$\hat{\lambda}^{raw}(x, \boldsymbol{\theta}) = N(x, \boldsymbol{\theta}) \Big/ \int_{u=0}^{1} E(x, u; \boldsymbol{\theta})\, F(t_{\theta_8} - \theta_5 + 1 - u; x, \boldsymbol{\theta})\, du \tag{7.4}$$

whose standard error can be estimated as

$$\sqrt{N(x, \boldsymbol{\theta})} \Big/ \int_{u=0}^{1} E(x, u; \boldsymbol{\theta})\, F(t_{\theta_8} - \theta_5 + 1 - u; x, \boldsymbol{\theta})\, du. \tag{7.5}$$

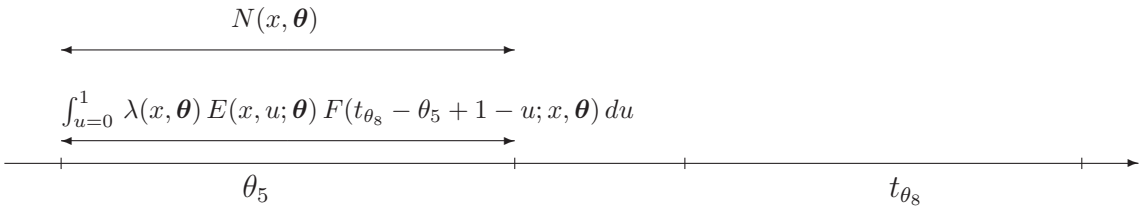


Figure 7.1: Observed and expected number of claims.

The integral for calculating the adjusted exposure given in (7.2) is evaluated using the repeated Simpson's Rule with four steps. The raw rates produced by this procedure can then be smoothed by specifying a model for the intensities and estimating the parameters by maximum likelihood. A discussion of smoothing methodology for mortality rates can be found in Forfar *et al.* (1988). Following that work, under the Poisson model given in (7.1), we regress the intensity on covariates of interest (defined in Table 7.1) for smoothing the crude diagnosis inception rates, $\lambda(x, \boldsymbol{\theta})$, in the following way

$$\lambda(x, \boldsymbol{\theta}) = g_r(x) + \exp\left(f_s(x) + \boldsymbol{\beta}\boldsymbol{\theta}\right), \qquad r, s = 0, 1, \ldots \tag{7.6}$$

where $g_r(x) = \sum_{i=1}^{r} \kappa_i\, x^{i-1}$, $f_s(x) = \sum_{j=1}^{s} \delta_j\, x^{j-1}$, x is standardised age last birthday, $\boldsymbol{\kappa}$ and $\boldsymbol{\delta}$ are vectors of coefficients for age and $\boldsymbol{\beta}$ is a vector of coefficients for the

covariates given in Table 7.1. Then the function given in (7.6) can be written explicitly as

$$\lambda(x, \boldsymbol{\theta}) = \sum_{i=1}^{r} \kappa_i \, x^{i-1} + \exp\left(\sum_{j=1}^{s} \delta_j \, x^{j-1} + \boldsymbol{\beta}\boldsymbol{\theta} \right). \tag{7.7}$$

This model will be used in this chapter for the all-causes inception rates and in Chapter 8 for the cause-specific inception rates. Note that with $r = 0$, the formula reduces to a GLM with Poisson errors and log-link function. When $r > 0$, we estimate the model coefficients appearing in (7.7) by maximising the log-likelihood function using a Newton-Raphson iterative method. Once the $\boldsymbol{\delta}$ and $\boldsymbol{\kappa}$ coefficients are estimated, fitted inception rates i.e. smoothed rates, $\hat{\lambda}(x, \boldsymbol{\theta})$, or fitted number of claims, $\hat{N}(x, \boldsymbol{\theta})$, can be calculated. A similar model is used by Richards (2008), where he applies Cox-type features to each parameter in his baseline model individually. Later in this chapter and in Chapter 8 when modelling individual causes, it will be seen that in all cases except one the best fit is obtained with $r = 0$. The one exception is for the cause-specific inception rates relating to death, where a model with $r = 1$ gives the best fit. In this case, the constant term, $g_0(x)$, presumably plays the same role as the constant A in Makeham's formula for the force of mortality: capturing deaths due to accidents, which are not related to age or any of the covariates.

Variable selection

We use the following procedure to determine the best model:

1. Variable selection on the covariate vector $\boldsymbol{\theta}$ is performed separately for each fixed order s of the f function, starting from the first order polynomial, $f_2(x)$.

2. Fixing the order s and the covariates which are found important in the model, we investigate whether adding a polynomial $g_r(x)$ to the model improves the model fit. A polynomial is introduced in the model starting from order 0, $g_1(x) = 1$. If the fit is better, then we increase the order of r by 1 and continue searching; if it is not, then we no longer investigate for higher order models. In previous analyses using a subset of these covariates at earlier stages of this research, we were able to select covariates under different $g_r(x)$ functions together with a given $f_s(x)$ function. Although we have not been able to verify for this set

of covariates (due to computing issues because of the size of the data), we would like to mention that the selected covariates under the $f_s(x)$ model were not different from those selected after introducing the linear polynomial, $g_r(x)$, using the reduced set of covariates.

The significance of the year of diagnosis is being investigated in variable selection through the covariate θ_5. When θ_5 is not part of the model, we add up the number of claims and exposures over the diagnosis year.

7.3 All-cause rates without restriction

In this section, all the covariates given in Table 7.1 are incorporated to estimate the intensity of diagnosis from all causes together with the interaction term between smoker status and age. The reason we searched for this interaction is that CMI WP 43 (2010) indicates that there might be an interaction between these two variables (see page 25). Considering that they are not using exactly the same data set, in Figure 7.2 we plot male smoker crude rates against male non-smoker crude rates with our data set. The graph we have is similar to the one in CMI WP 43 (2010). The 'U-shape' can be interpreted as the limited effect of smoking on health for younger ages and various other risk factors independent of smoking for older ages.

Table 7.2 shows the log-likelihood and BIC values of the selected models under different orders of $g_r(x)$ and $f_s(x)$ polynomials. According to this, including the $g_0(x)$, $f_2(x)$ functions with the smoker status (θ_3), policy duration (θ_7), office (θ_8) and age-smoker status interaction ($x \times \theta_3$) covariates is found to provide the best model. Although gender is generally accepted as an important factor in diagnosis inception rates, we think the reason for this covariate being dropped from the best model is the neutralising effect of different individual causes for different genders.

The summary of the estimates of the best model is given in Table 7.3. Model coefficients suggest that age has a significant positive effect on the diagnosis inception rates (Note here that age is standardised by subtracting the mean (39.75) and dividing by the standard deviation (11.21)). All-cause rates are higher for smokers compared to non–smokers. Up to curtate policy duration 4 years, the claim diagnosis rates are in-

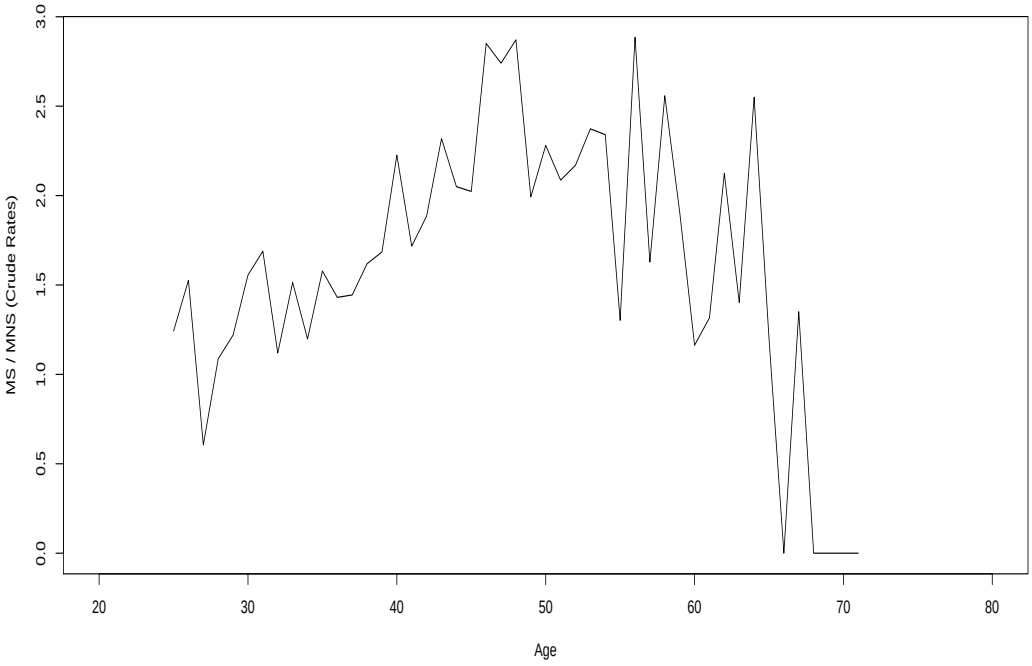


Figure 7.2: Male smoker crude inception rates divided by male non-smoker crude inception rates.

Table 7.2: Selected covariates, log-likelihood values and BIC from fitting different $g_r(x), f_s(x)$ polynomials.

Polynomials	Covariates	l	BIC
$g_0(x), f_2(x)$	$\theta_3, \theta_7, \theta_8, x \times \theta_3$	-67373.3	135025
$g_1(x), f_2(x)$	$\theta_3, \theta_7, \theta_8, x \times \theta_3$	-67372.7	135038
$g_0(x), f_3(x)$	$\theta_3, \theta_7, \theta_8, x \times \theta_3, x^2 \times \theta_3$	-67365.6	135037

creasing. This is what we expect, if there is no anti-selection. However, the coefficient for policy duration 4 is less than for policy duration 3. This issue is also mentioned by CMI in WP 43 (2010) and the problem is handled by putting a constraint on policy duration stating that the rates can not decrease with increasing duration unless there is anti-selection. The combined effect of two competing forces of selection, i.e. positive selection from health checks at the beginning of the policy and anti-selection arising from insufficient underwriting (e.g. non-revelation of medical history), should be investigated before forcing the data to overfit. In the CMI's WP 43 (2010), the combined effect is said to be not obvious but the constraint is applied in their work assuming that there is a smooth progression of rates by policy duration. In this study we are not forcing the rates to be higher. The other significant covariate is the Office

variable. Among 13 offices, Office 7 has the lowest rates and Office 11 has the highest rates. The coefficient of Office 8 gives the median effect. In the model, the age-smoker interaction term is found to be significant with a positive effect on the inception rates. Model fit is assessed using Pearson's χ^2 and the p-value is found to be 0.0174.

Table 7.3: ML estimates of parameters under the best model for all-cause rates.

Parameter	Estimate	Std. Error	p-value
$\delta_{intercept}$	-6.6180	0.0224	$< 2 \times (10^{-16})$
δ_{zage}	0.8954	0.0109	$< 2 \times (10^{-16})$
β_{smoker}	0.3957	0.0185	$< 2 \times (10^{-16})$
$\beta_{poldur0}$	-0.1375	0.0166	$< 2 \times (10^{-16})$
$\beta_{poldur1}$	0.0080	0.0159	0.6159
$\beta_{poldur2}$	0.0505	0.0167	0.0025
$\beta_{poldur3}$	0.0681	0.0188	0.0003
$\beta_{poldur4}$	-0.0441	0.0233	0.0586
$\beta_{poldur5+}$	0.0550	0.0174	0.0016
$\beta_{office1}$	0.0970	0.0300	0.0012
$\beta_{office2}$	0.1583	0.0274	$< 8 \times (10^{-9})$
$\beta_{office3}$	-0.0923	0.0777	0.2348
$\beta_{office4}$	-0.0065	0.0583	0.9116
$\beta_{office5}$	0.0653	0.0494	0.1862
$\beta_{office6}$	-0.0302	0.0900	0.7374
$\beta_{office7}$	-0.5106	0.2070	0.0136
$\beta_{office8}$	-0.00005	0.0291	0.9987
$\beta_{office9}$	0.1150	0.0325	0.0004
$\beta_{office10}$	-0.0608	0.0453	0.1793
$\beta_{office11}$	0.3027	0.0253	$< 2 \times (10^{-16})$
$\beta_{office12}$	0.2182	0.0310	$< 2 \times (10^{-12})$
$\beta_{office13}$	-0.2562	0.0487	$< 2 \times (10^{-7})$
$\beta_{zage \times smoker}$	0.2044	0.0204	$< 2 \times (10^{-16})$

Monitoring the behaviour of the diagnosis inception rates especially at younger ages is difficult on an actual scale. Therefore we change to a log scale in order to see the details. Figures 7.3 to 7.8 show the inception rates against age on a log scale for non-smokers – policy durations 0 to 5+ (NS0 to NS5+) and smokers – policy durations 0 to 5+ (S0 to S5+), respectively. Note that y-axis is given in original scale. In each graph, crude rates are shown with a dark blue line together with their ± 2 standard errors (dotted blue lines). In these graphs, the crude rates (solid dark blue line) and the smoothed rates (red line) are weighted averages for offices. Since our model includes office as a covariate, the modelled inception rates will depend on office as well as smoker status and policy duration. To obtain weighted smoothed

rates for offices we use (7.8).

$$\frac{\sum_{i=1}^{13} \hat{\lambda}(x;\theta_{8,i};\theta_{\setminus\theta_8}) E^*(x;\theta_{8,i};\theta_{\setminus\theta_8})}{\sum_{i=1}^{13} E^*(x;\theta_{8,i};\theta_{\setminus\theta_8})} \tag{7.8}$$

Here $\theta_{8,i}$ denotes Office$_i$ for $i = 1, \ldots, 13$ and $\theta_{\setminus\theta_8}$ denotes the other (fixed) characteristics except office. In other words, to find weighted smoothed rates for offices (red lines), we first obtain the inception rates from the best model (these inception rates are office-specific rates) and then these office-specific inception rates are weighted with office-specific exposures with the same characteristics and divided into the sum of these exposures. If crude rates rather than modelled rates are used the weighted average for crude rates will be obtained.

To see the effects of individual offices, we include three offices' rates in the graphs. These are Office 7, Office 11 and Office 8 which have the highest, the lowest and the median coefficients in our model, respectively.

In most of the cases crude rates are provided until around age 65 as there are very few data beyond this age. To obtain the weighted smoothed inception rates for offices the lack of exposure after age 65 caused lack of smoothness on the smoothed rates after that age. Therefore, we fixed the exposure for older ages by the exposure at age 65. This problem also arises for younger ages as we have hardly any data for ages below (about) 20. On the other hand, we have not fixed the exposure for these younger ages. The lack of smoothness for these ages can be seen clearly in Figures 7.5 and 7.8 as long policy commitment is not very common at younger ages.

Note that in all figures weighted smoothed rates start closer to the median-effect office (Office 8) and approach the office with highest rates (Office 11) with increasing age. The reason for this is that its weight is higher for older ages as for this age range most of the data are coming from this office whereas for younger ages data are coming from all offices. The smoothed inception rates for the lowest rates office (Office 7) mostly lie outside the lower bound of the confidence intervals for weighted crude rates for offices. The reason for this is the small business volume of this office.

Overall, for all of the combinations of characteristics considered, smoothed rates obtained from our analysis lie within the confidence interval of the crude rates.

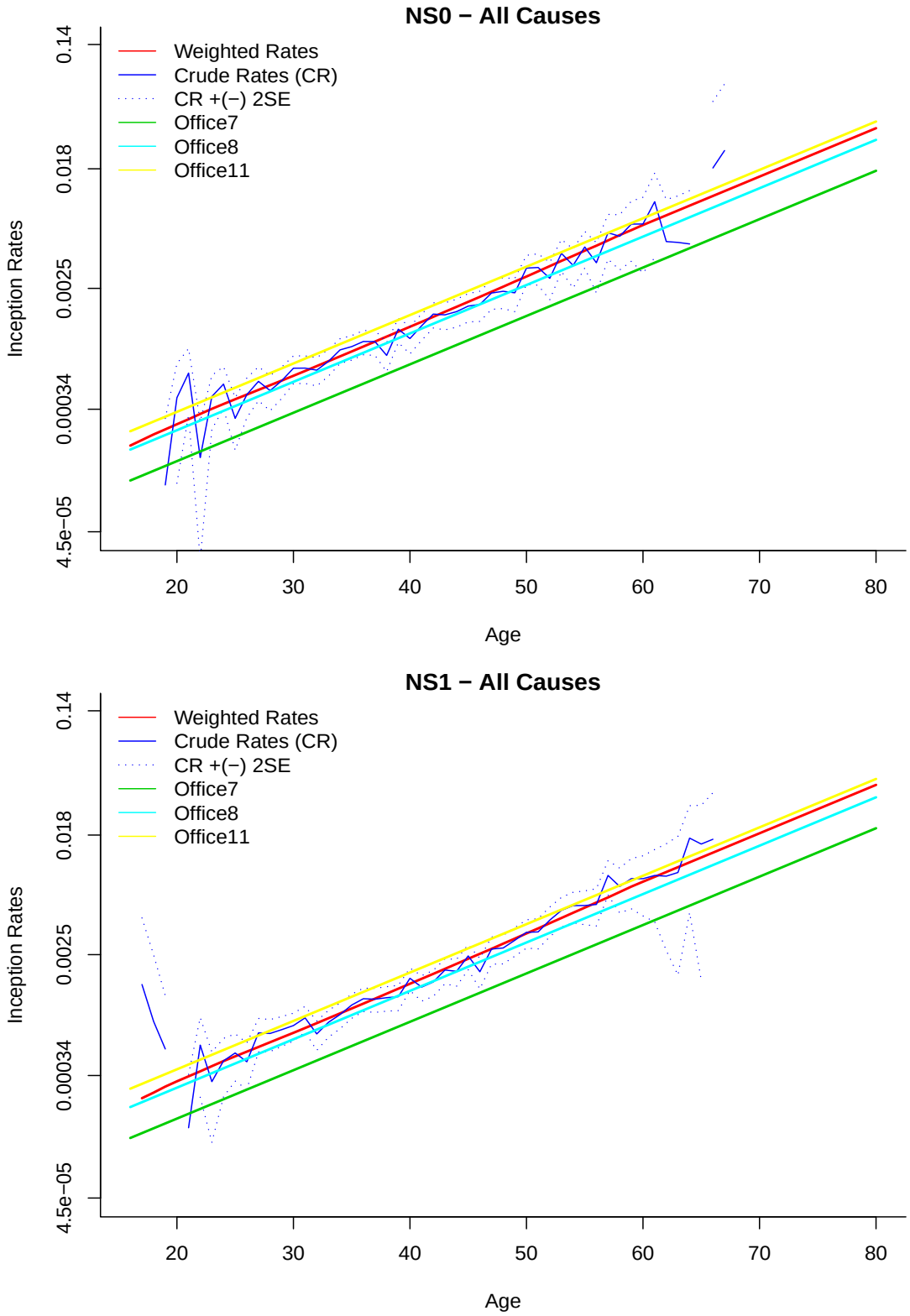


Figure 7.3: Graphs of diagnosis inception rates for non–smokers and durations 0 & 1.

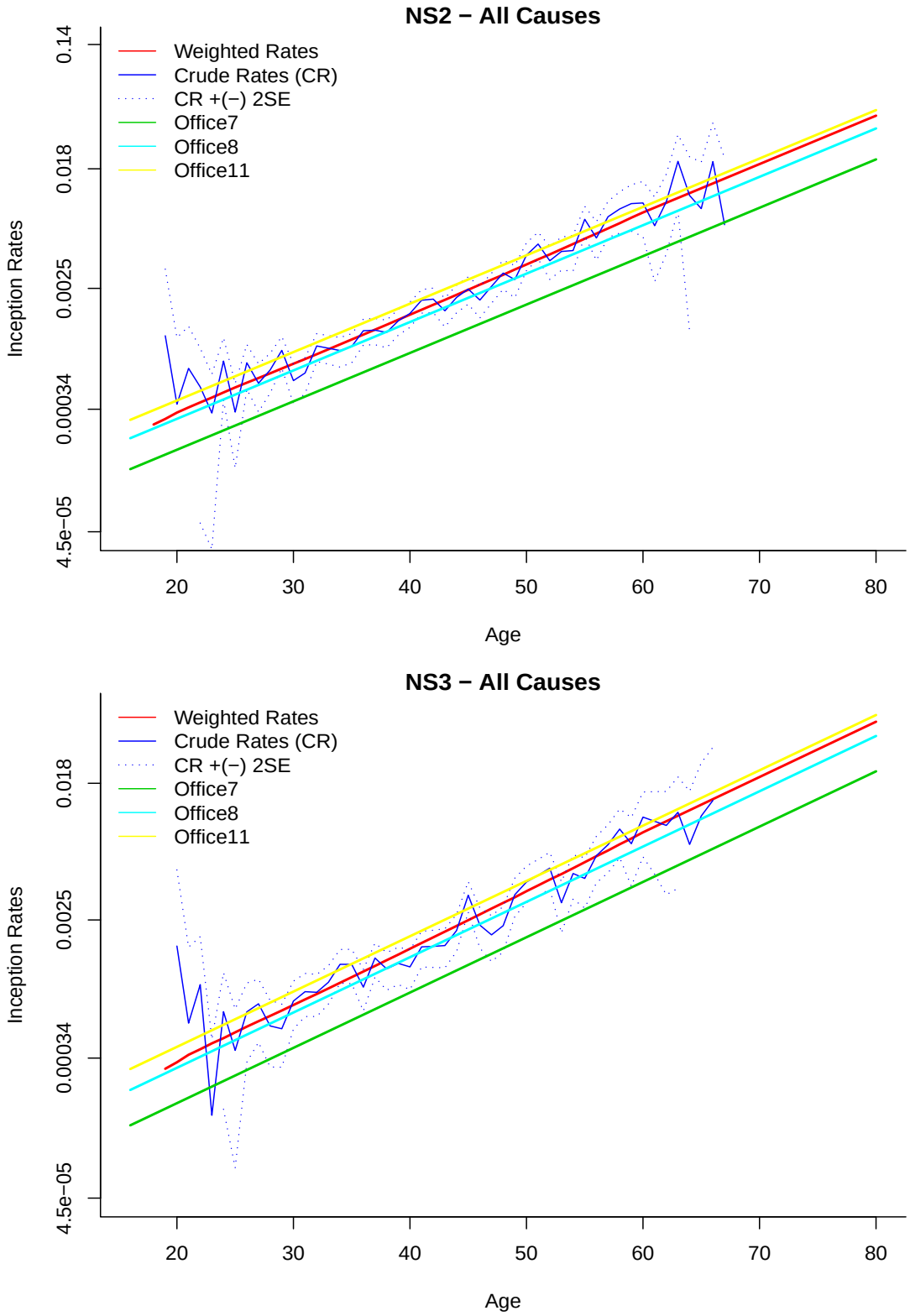


Figure 7.4: Graphs of diagnosis inception rates for non–smokers and durations 2 & 3.

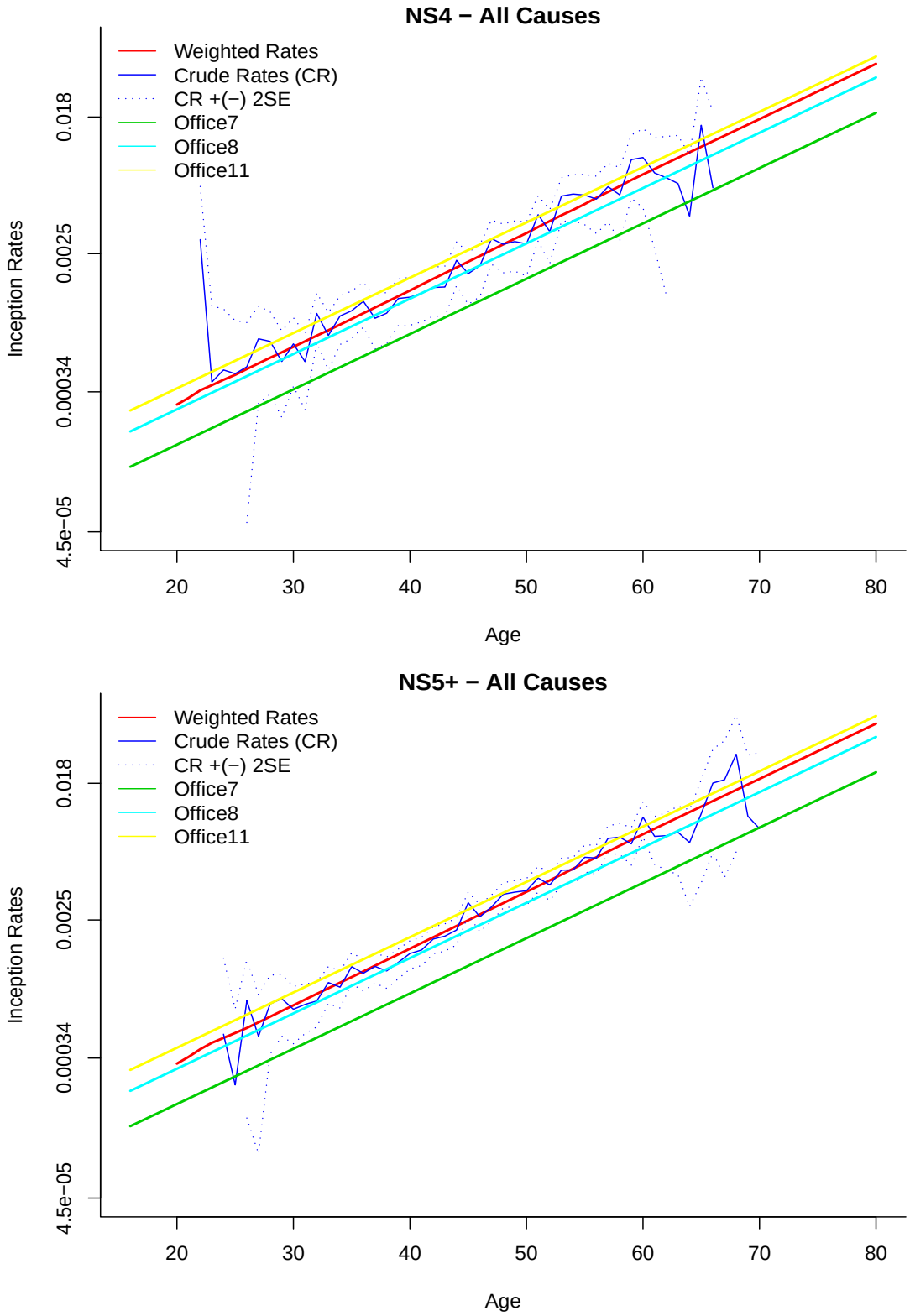


Figure 7.5: Graphs of diagnosis inception rates for non–smokers and durations 4 & 5+.

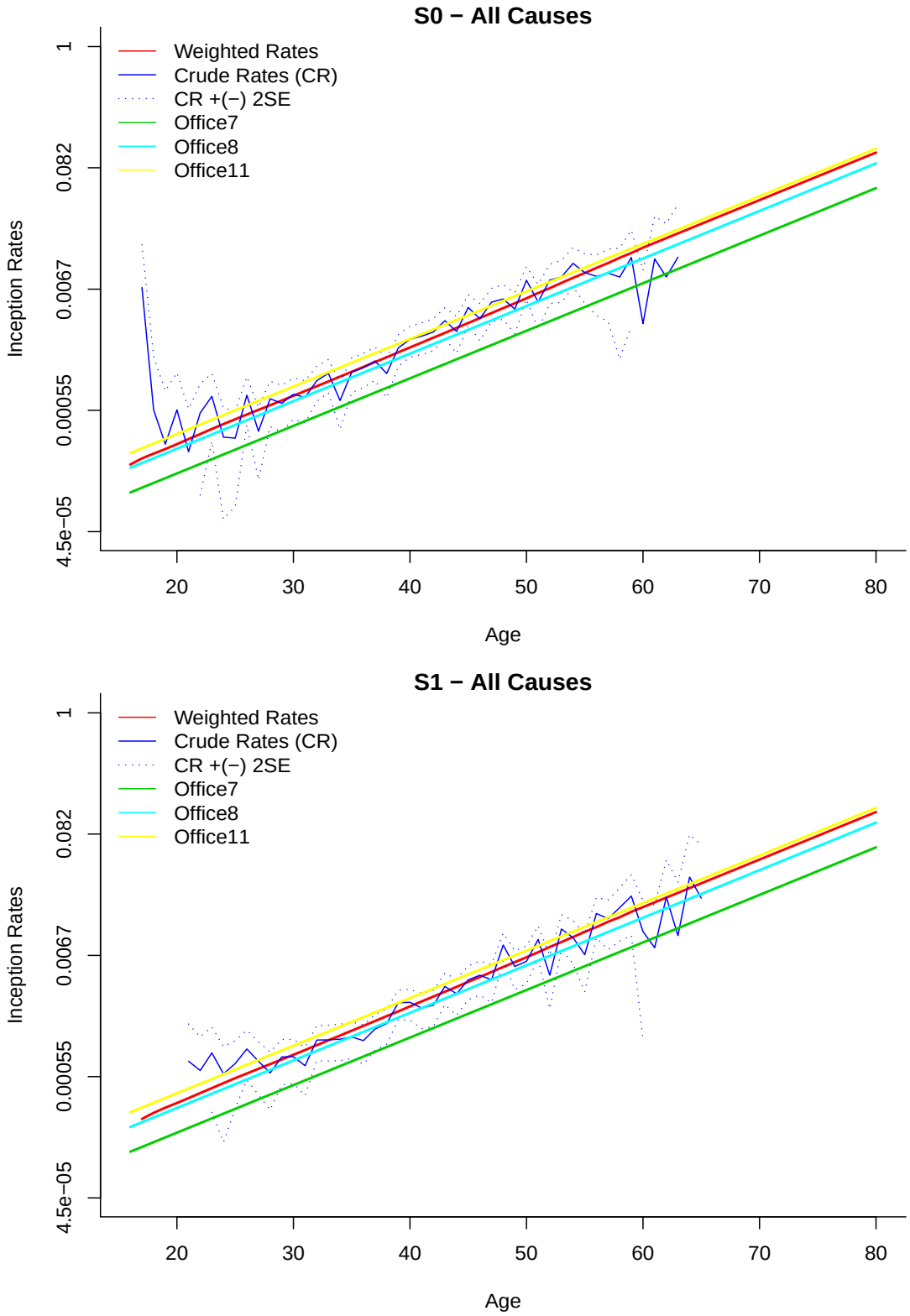


Figure 7.6: Graphs of diagnosis inception rates for smokers and durations 0 & 1.

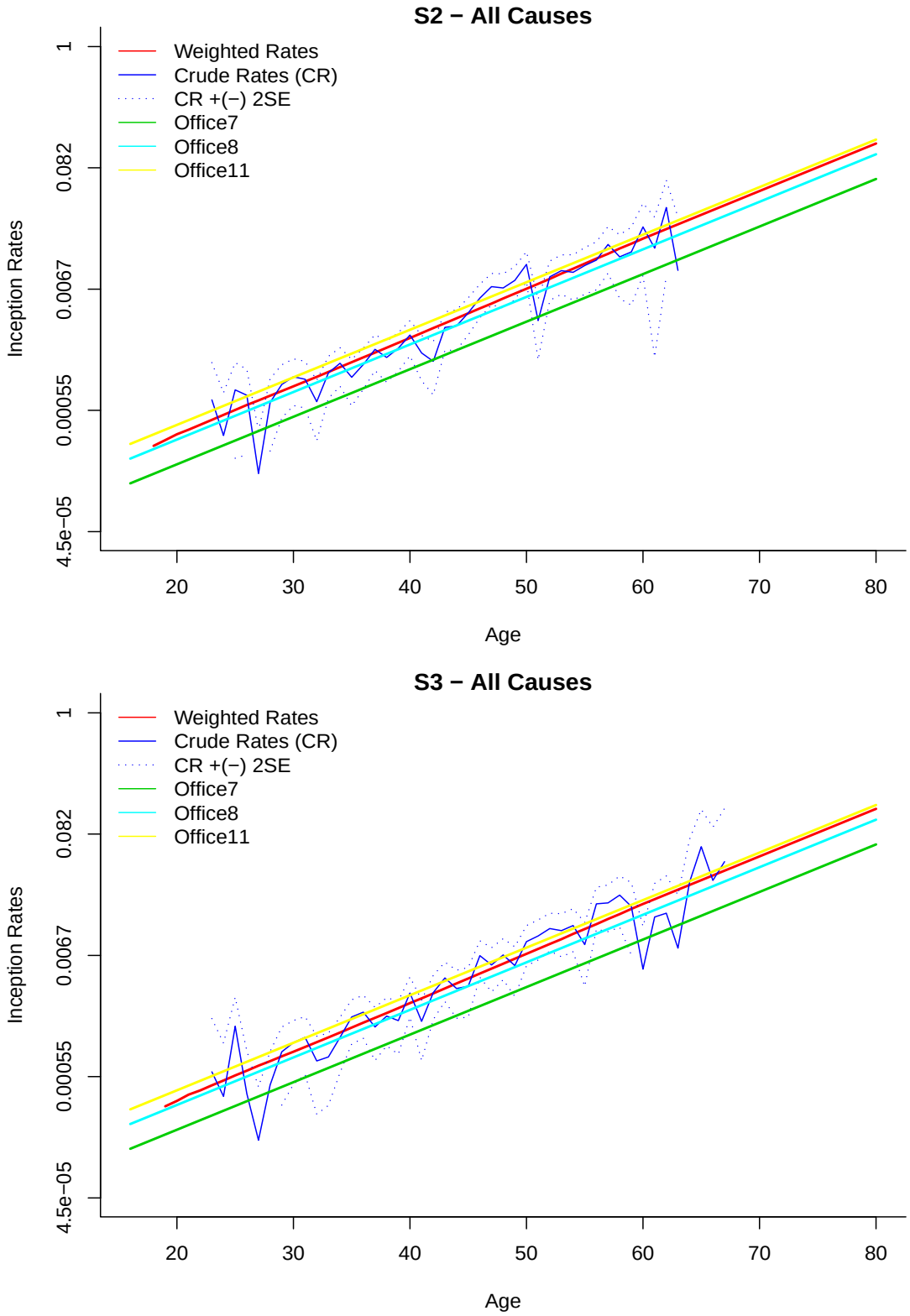


Figure 7.7: Graphs of diagnosis inception rates for smokers and durations 2 & 3.

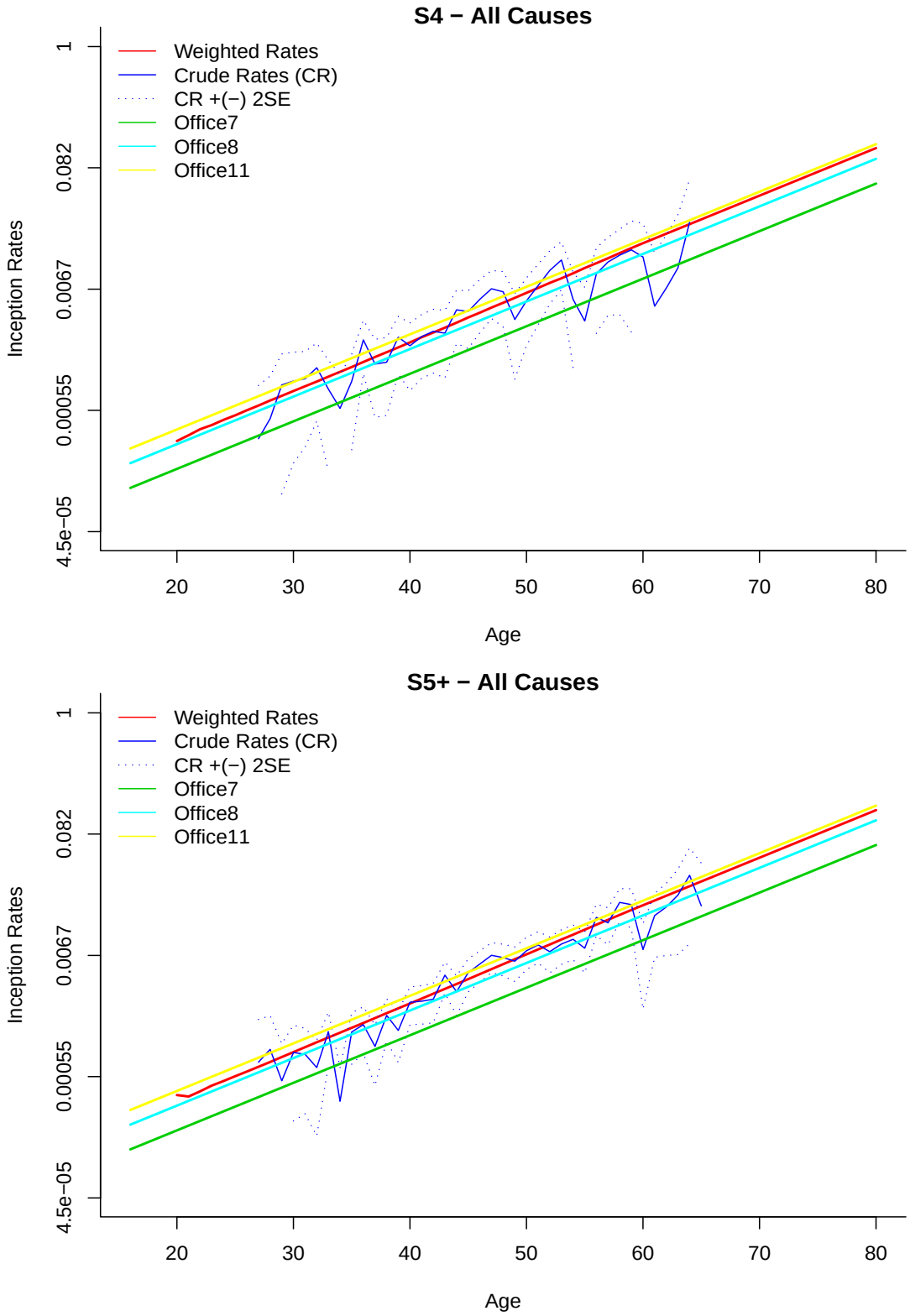


Figure 7.8: Graphs of diagnosis inception rates for smokers and durations 4 & 5+.

Modelled smoker rates against non–smoker rates are shown in Figure 7.9. Because of the age - smoker interaction in the model the two lines cross below age 20. From that age on, smoker rates are always higher than non-smoker rates, as expected. For the ages below 20 the rates should be adjusted either by increasing the smoker rates to the level of non-smokers rates or decreasing the non-smokers rates to the level of smokers rate.

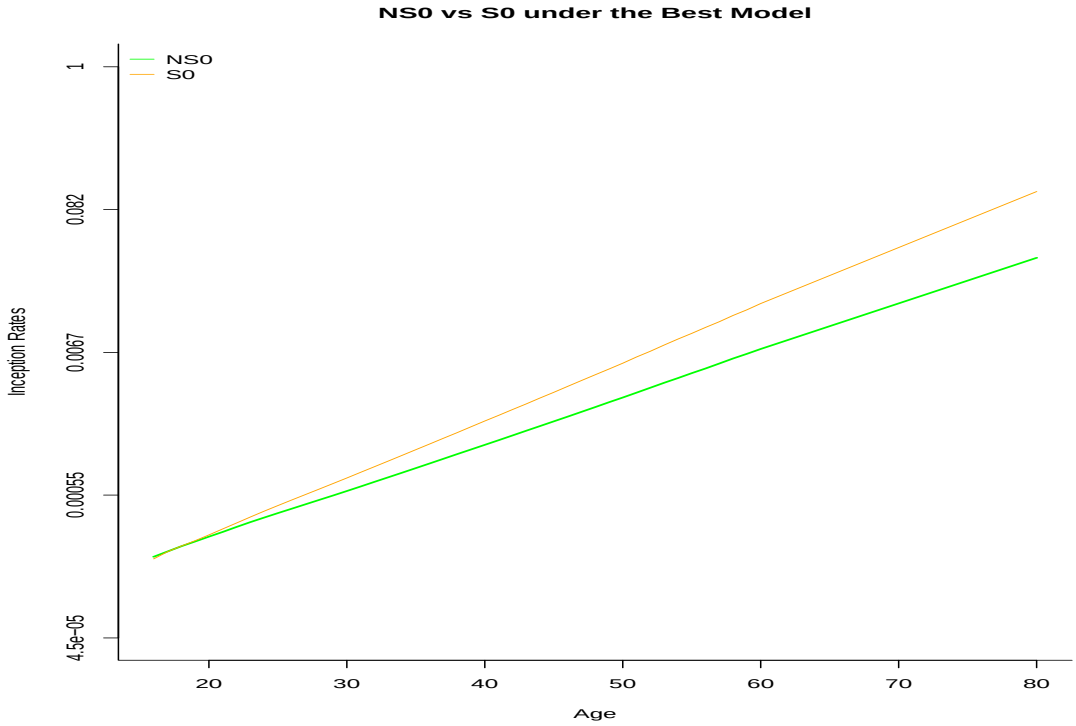


Figure 7.9: Comparison of diagnosis inception rates for non–smokers vs smokers under the best model for policy duration 0.

7.4 All-cause rates including the CMI variables

In this section we compare the estimated rates from our analysis with the the claim inception rates calculated by the CMI. We were kindly provided with these rates by the CMI. Although the rates are provided from age 20 to 80, the CMI mentions that outside the age range 25 - 65 the rates are only indicative due to lack of data. These rates are presented in CMI WP 43 (2010). The rates are based on claims settled between 1999 and 2004. Although this data set and our data set overlap considerably, there are some differences between them mainly arising from different contributing offices. In WP 43 (2010), all-cause rates are produced for each combination of sex, smoker status and policy duration without examining the statistical significance of these variables in the modelling. Also, the rates are given for full accelerated policies only. In other words, the CMI variables are

sex

smoker status

policy duration

benefit type (for full accelerated policies).

The definitions of these variables are the same as the ones given in Table 7.1. However note that the age definition CMI uses is the 'age nearest' whereas our rates are produced under the 'age last birthday' definition. This difference will shift each age range for the inception rates by half a year.

Model 1:

To compare the CMI's rates with these obtained here, we find the best model which includes the CMI variables. To clarify, we force these four variables to stay in the model and search for other possible covariates when they are in the model. Office is the only variable added to these four covariates after a stepwise search. So the model includes

sex

smoker status

policy duration

benefit type (for full accelerated policies) and

office.

Table 7.4 shows selected models under different orders of age polynomials, together with their log-likelihood and BIC values. Among these models, the one including the $g_0(x), f_2(x)$ functions with sex (θ_1), benefit type (θ_2), smoker status (θ_3), policy duration (θ_7), office (θ_8) and age - smoker status interaction ($x \times \theta_3$) covariates is found to be the best. The estimated parameters under this model are summarised in Table 7.5. The difference between this model and the best model presented in Section 7.3 is that this model involves sex and benefit type as covariates. This allows us to find out how the inception rates are changing across the genders. According to the coefficient, males have higher claim diagnosis rates than females (F is the base category, see Table 7.1). In addition to that, stand alone policies have lower inception rates than full accelerated policies (FA is the base category).

Table 7.4: Selected covariates, log-likelihood values and BIC from fitting different $g_r(x), f_s(x)$ polynomials.

Polynomials	Covariates	l	BIC
$g_0(x), f_2(x)$	$\theta_1, \theta_2, \theta_3, \theta_7, \theta_8, x \times \theta_3$	-67363.6	135033
$g_1(x), f_2(x)$	$\theta_1, \theta_2, \theta_3, \theta_7, \theta_8, x \times \theta_3$	-67363.0	135045
$g_0(x), f_3(x)$	$\theta_1, \theta_2, \theta_3, \theta_7, \theta_8, x \times \theta_3, x^2 \times \theta_3$	-67356.0	135044

Table 7.5: ML estimates of parameters under the best model which includes the CMI variables for all-cause rates.

Parameter	Estimate	Std. Error	p-value
$\delta_{intercept}$	-6.6360	0.0243	$< 2 \times (10^{-16})$
δ_{zage}	0.8937	0.0110	$< 2 \times (10^{-16})$
β_{sex}	0.0543	0.0152	0.0004
β_{smoker}	0.3912	0.0185	$< 2(10^{-16})$
$\beta_{bentype}$	-0.0631	0.0242	0.0091
$\beta_{poldur0}$	-0.1392	0.0166	$< 2 \times (10^{-16})$
$\beta_{poldur1}$	0.0069	0.0159	0.6667
$\beta_{poldur2}$	0.0501	0.0167	0.0027
$\beta_{poldur3}$	0.0686	0.0188	0.0003
$\beta_{poldur4}$	-0.0425	0.0233	0.0685
$\beta_{poldur5+}$	0.0562	0.0174	0.0012
$\beta_{office1}$	0.0892	0.0302	0.0031
$\beta_{office2}$	0.1567	0.0275	$< 2 \times (10^{-8})$
$\beta_{office3}$	-0.0935	0.0777	0.2290
$\beta_{office4}$	-0.0074	0.0583	0.8989
$\beta_{office5}$	0.0739	0.0495	0.1351
$\beta_{office6}$	-0.0149	0.0902	0.8689
$\beta_{office7}$	-0.5139	0.2070	0.0130
$\beta_{office8}$	0.0031	0.0291	0.9149
$\beta_{office9}$	0.1122	0.0325	0.0006
$\beta_{office10}$	-0.0521	0.0453	0.2502
$\beta_{office11}$	0.2949	0.0255	$< 2 \times (10^{-11})$
$\beta_{office12}$	0.2104	0.0311	$< 2 \times (10^{-16})$
$\beta_{office13}$	-0.2584	0.0488	$< 2 \times (10^{-7})$
$\beta_{zage \times smoker}$	0.2058	0.0204	$< 2 \times (10^{-16})$

Model 2:

We also fit a model using the four CMI variables only (i.e. sex, smoker status, policy duration and benefit type). This means that office is excluded from the above model. The log-likelihood and BIC values of the models under different age functions are given in Table 7.6. As in the other cases, $g_0(x), f_2(x)$ is found to give the best model. The estimated parameters under this model are summarised in Table 7.7.

Table 7.6: Selected covariates, log-likelihood values and BIC from fitting different $g_r(x), f_s(x)$ polynomials.

Polynomials	Covariates	l	BIC
$g_0(x), f_2(x)$	$\theta_1, \theta_2, \theta_3, \theta_7, x \times \theta_3$	-67508.1	135162
$g_1(x), f_2(x)$	$\theta_1, \theta_2, \theta_3, \theta_7, x \times \theta_3$	-67507.4	135174
$g_0(x), f_3(x)$	$\theta_1, \theta_2, \theta_3, \theta_7, x \times \theta_3, x^2 \times \theta_3$	-67498.9	135171

Table 7.7: ML estimates of parameters under the model with the CMI variables for all-cause rates.

Parameter	Estimate	Std. Error	p-value
$\delta_{intercept}$	-6.5051	0.0130	$< 2 \times (10^{-16})$
δ_{zage}	0.9224	0.0107	$< 2 \times (10^{-16})$
β_{sex}	0.0564	0.0152	0.0002
β_{smoker}	0.4023	0.0185	$< 2 \times (10^{-16})$
$\beta_{bentype}$	-0.1238	0.0235	$< 2 \times (10^{-7})$
$\beta_{poldur0}$	-0.1479	0.0162	$< 2 \times (10^{-16})$
$\beta_{poldur1}$	0.0043	0.0156	0.7831
$\beta_{poldur2}$	0.0498	0.0165	0.0026
$\beta_{poldur3}$	0.0707	0.0188	0.0002
$\beta_{poldur4}$	-0.0404	0.0233	0.0820
$\beta_{poldur5+}$	0.0636	0.0156	$< 5 \times (10^{-5})$
$\beta_{zage \times smoker}$	0.2112	0.0204	$< 2 \times (10^{-16})$

Figures 7.10 to 7.21 show the inception rates with respect to age on a log scale for combinations of sex, smoker status, and policy durations for full accelerated policies. Note that y-axis is given in original scale. The letters of the captions of the figures refer to sex (M or F), then the smoker status (NS or S) and then the benefit type (all of them are FA) and final numbers are the policy durations from 0 to 5+. In each graph, crude rates are shown with a dark blue line together with their ± 2 standard errors (dotted lines). These crude rates are the weighted averages for offices. The red line shows the smoothed rates from model 1 when the weighted average of the offices is taken (in the same way as in (7.8)). To show the smoothed rates for different offices, the lowest (Office 7), the highest (Office 11) and the median (Office 8) offices are represented on the graphs. Smoothed rates obtained by using model 2 are shown by a black line. This model excludes the Office covariate from the analysis. As can be seen from the graphs, the red lines and black lines are very close to each other.

The CMI rates are very close to the smoothed rates obtained from our analysis for the age range between 30 and 60. This is the range where we have most of the data. On the other hand when there is a big variability in the crude rates, the CMI rates have a better fit. This is because the CMI rates are essentially estimated by adjusting the crude rates by the help of base tables (e.g. CIBT02 (CI Trends Research Group, 2006)). See, for example, Figures 7.10 - 7.14 . In all of these graphs, for ages below 30, the CMI rates capture the crude rates better. Although the smoothed rates obtained from our analysis lie below the crude rates for males for this age range, they are still

within two standard errors.

Confidence intervals are wider for younger and older ages due to lack of data for these ages. When the number of observations is very small, the lower confidence limit cannot be provided. For example for the upper graph of Figure 7.10, the lower confidence limit could not be calculated for ages above 60 due to the small number of observed claims at that age with the other characteristics. Having less data at older ages for very short policy durations is not surprising. In a similar way, there are less data at younger ages for long policy durations, since long policy commitments for these ages are unusual (see e.g. Figures 7.12 , 7.15 , 7.18 and 7.21). In general, the confidence intervals are narrower for the non–smoker graphs compared to smokers as we have more data for non–smokers. On the upper graph of Figure 7.21 it is seen that almost none of the lower bounds could be given, again due to lack of data for female smokers with policy duration 4 years.

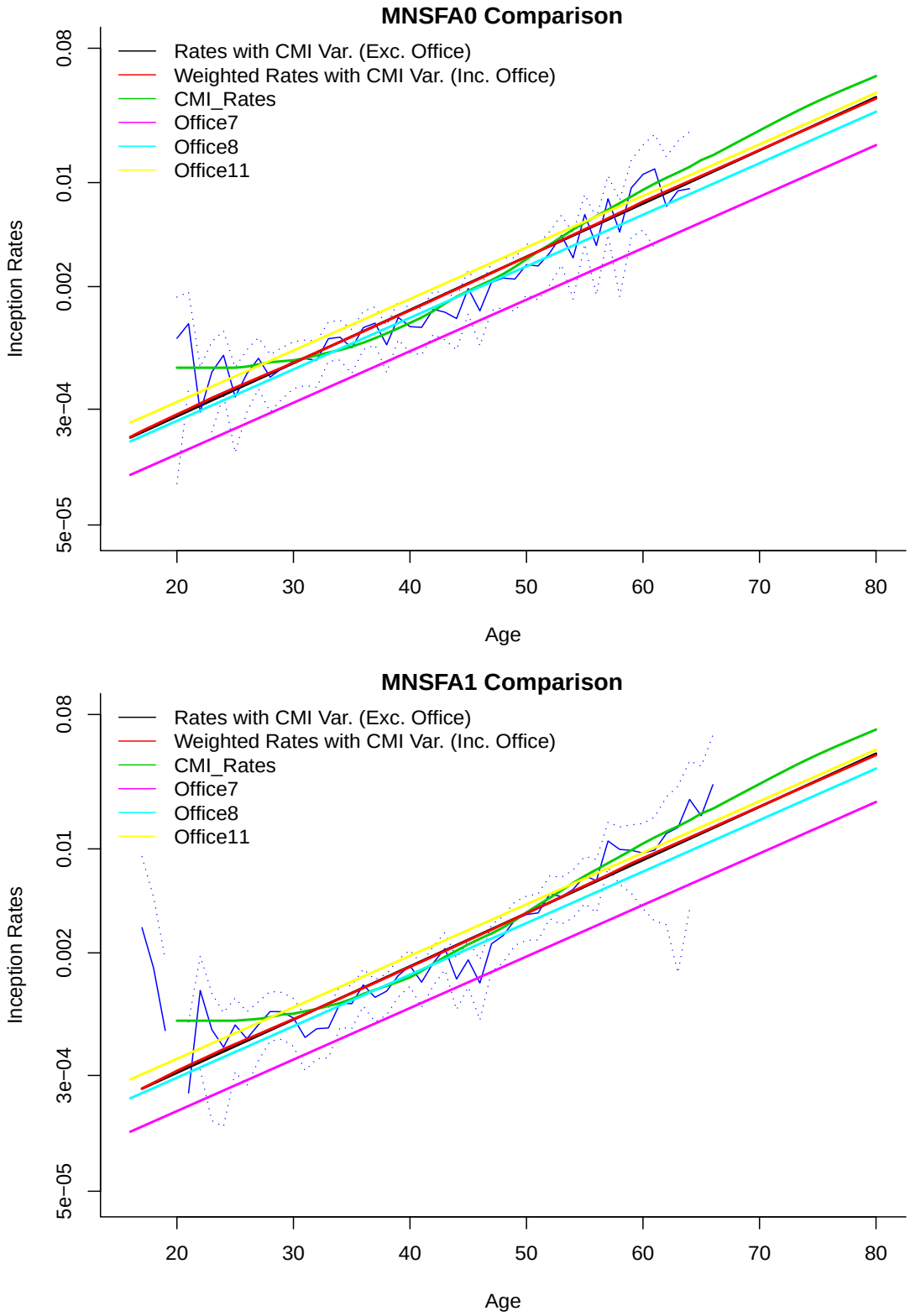
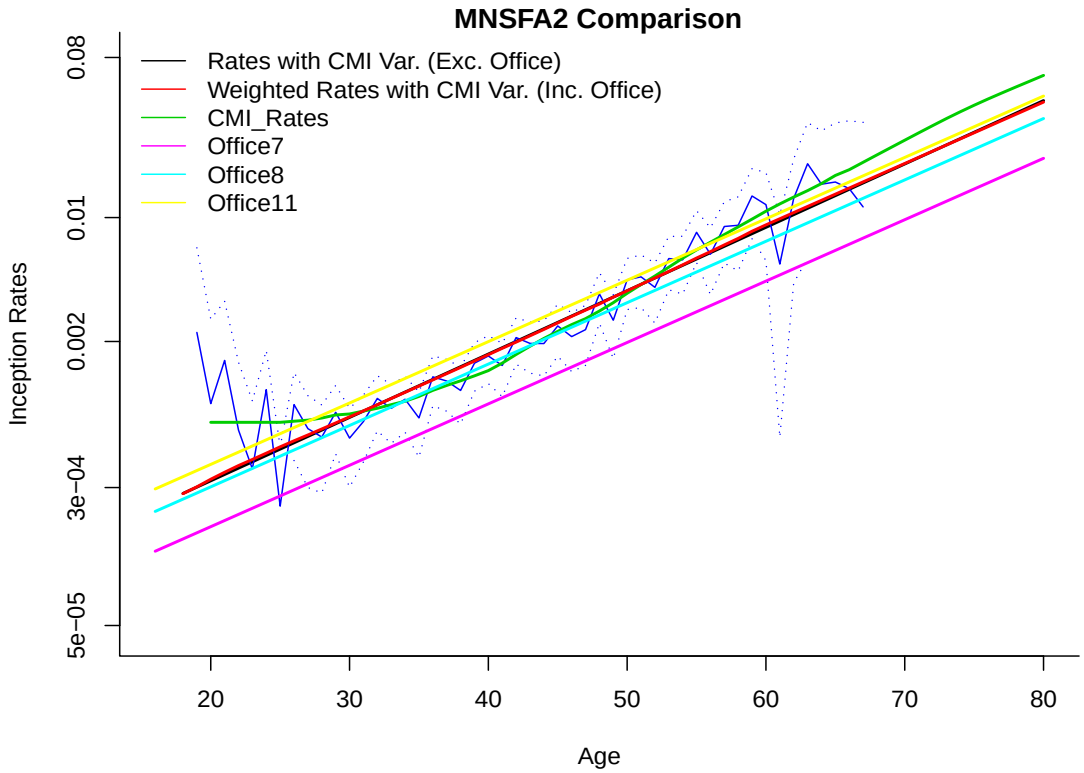


Figure 7.10: Graphs of diagnosis inception rates for males, non–smokers, full accelerated policies and durations 0 & 1.



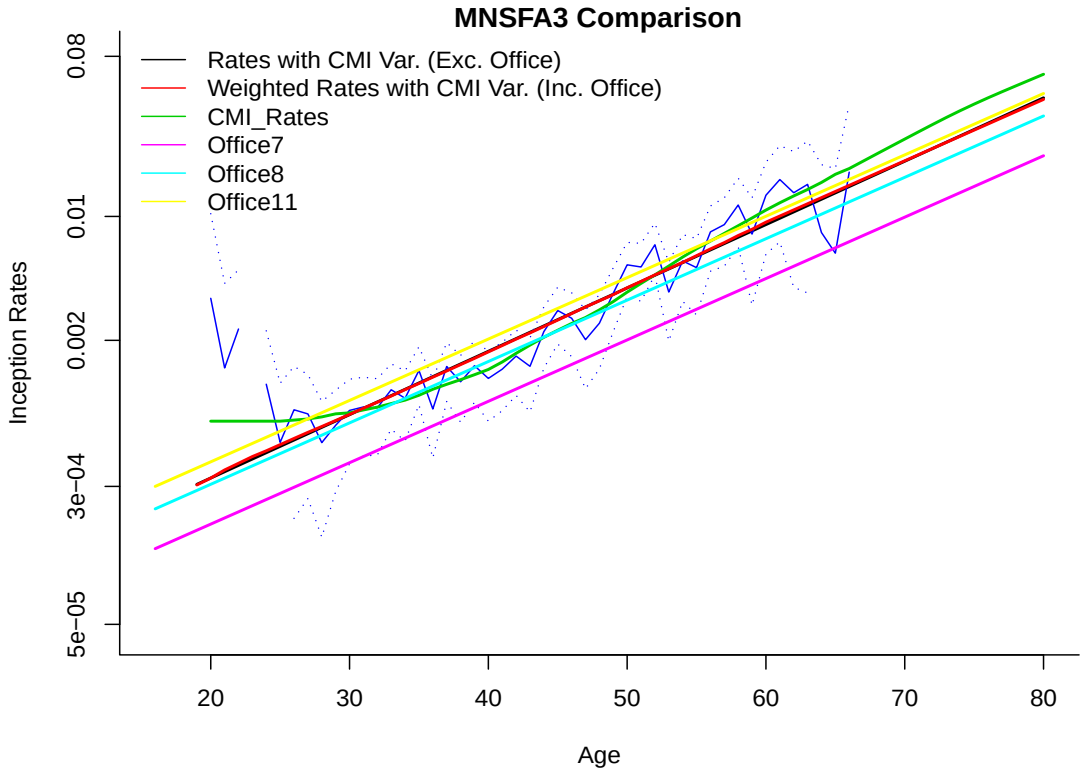


Figure 7.11: Graphs of diagnosis inception rates for males, non–smokers, full accelerated policies and durations 2 & 3.

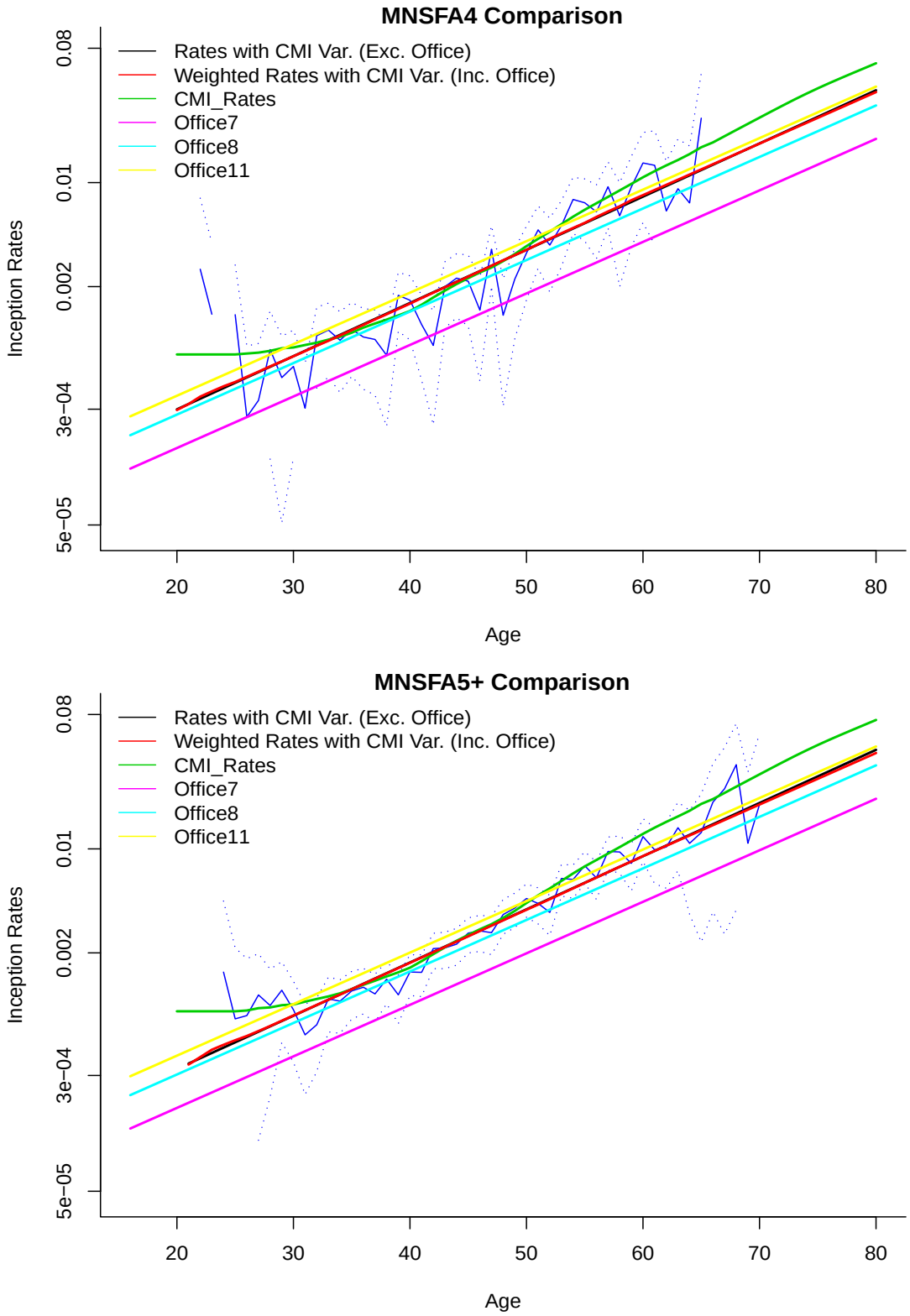


Figure 7.12: Graphs of diagnosis inception rates for males, non–smokers, full accelerated policies and durations 4 & 5+.

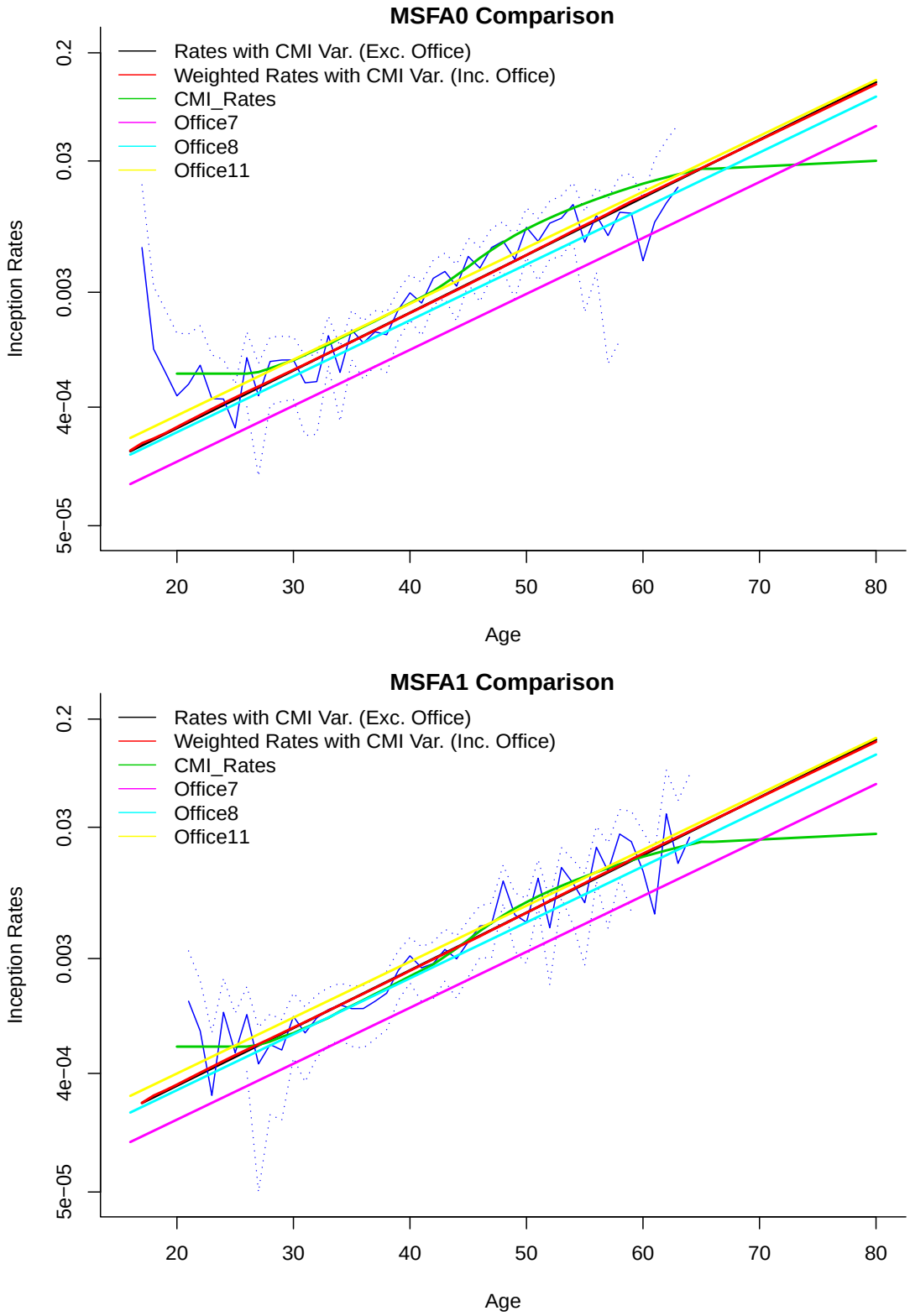


Figure 7.13: Graphs of diagnosis inception rates for males, smokers, full accelerated policies and durations 0 & 1.

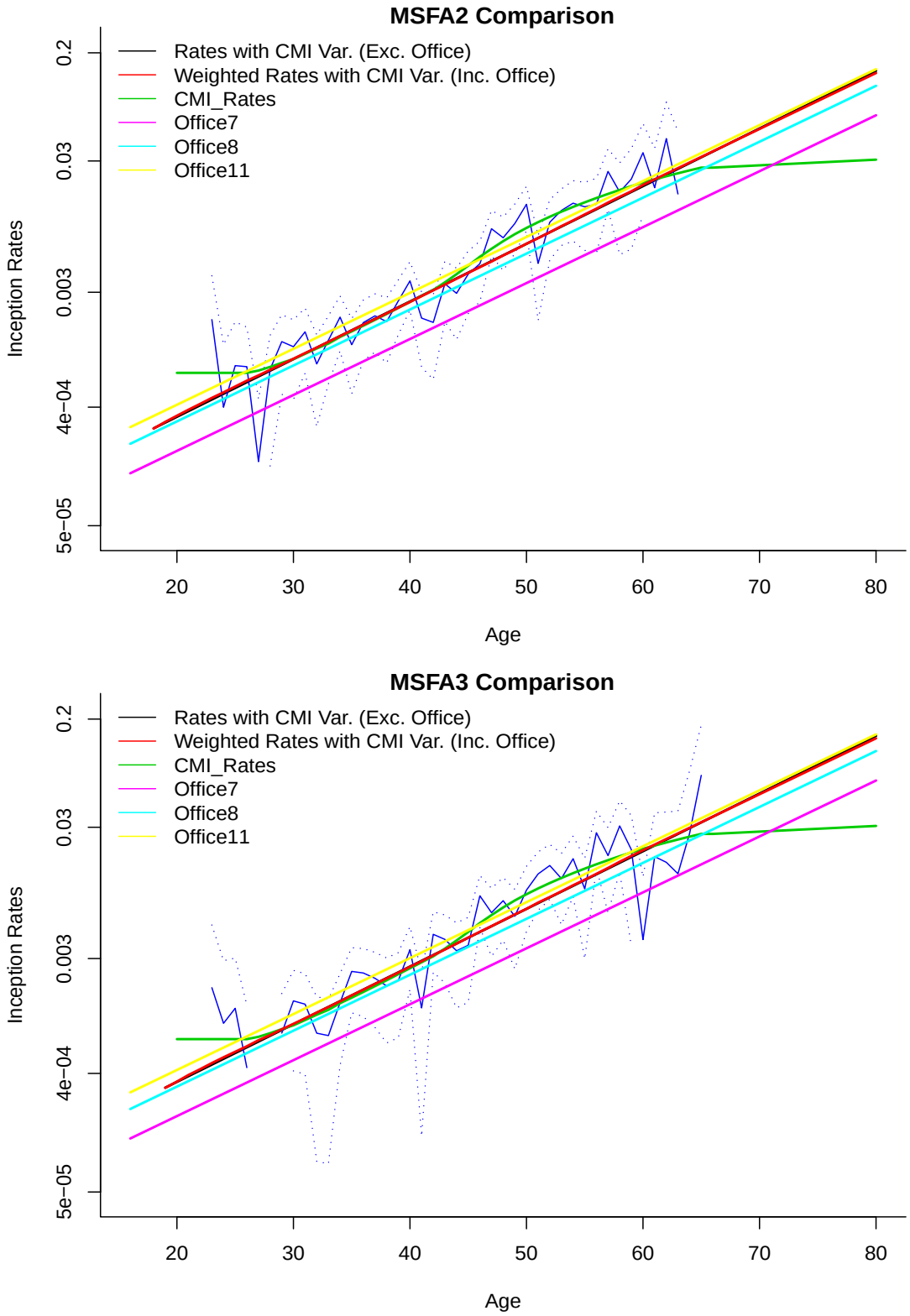


Figure 7.14: Graphs of diagnosis inception rates for males, smokers, full accelerated policies and durations 2 & 3.

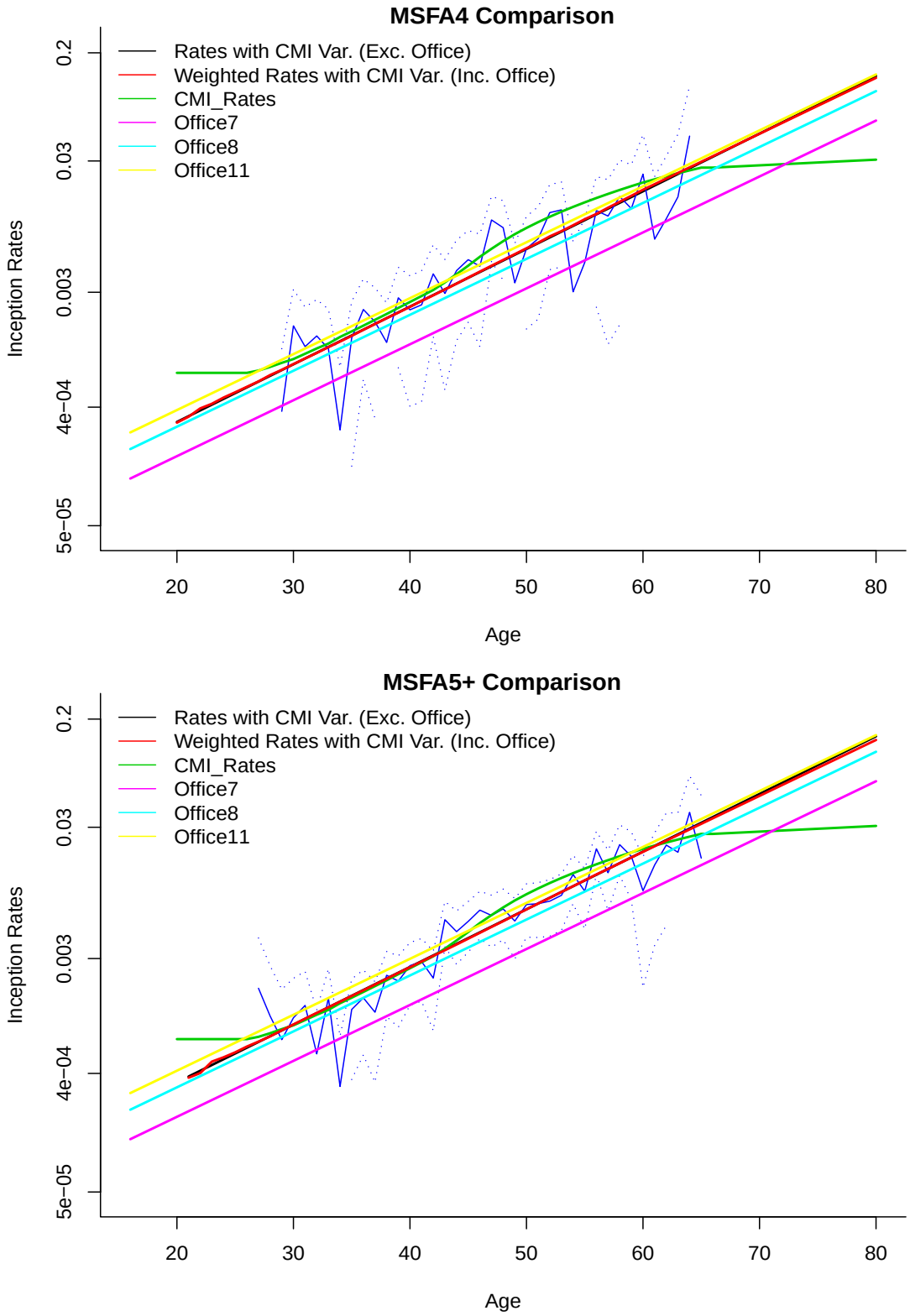


Figure 7.15: Graphs of diagnosis inception rates for males, smokers, full accelerated policies and durations 4 & 5+.

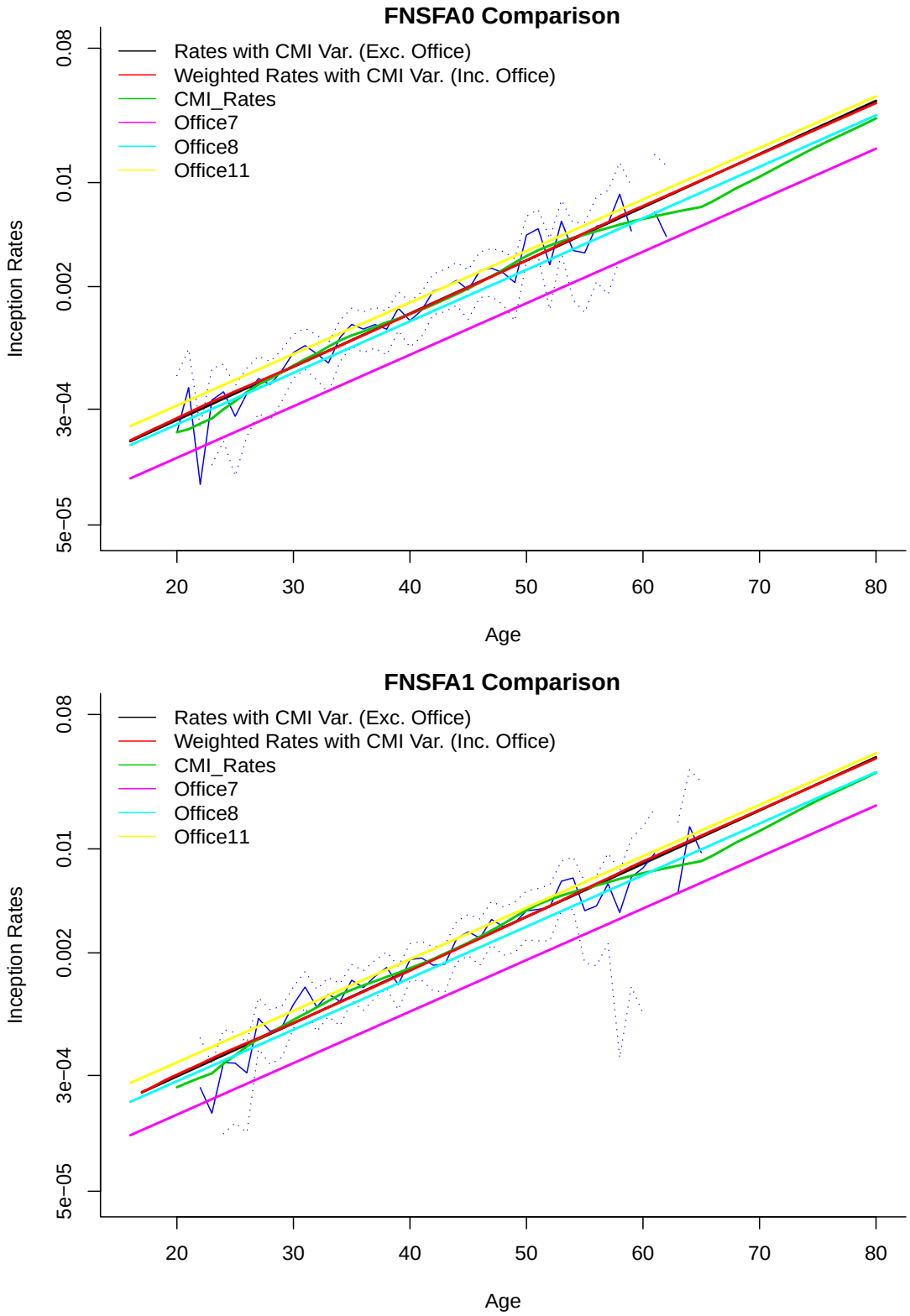
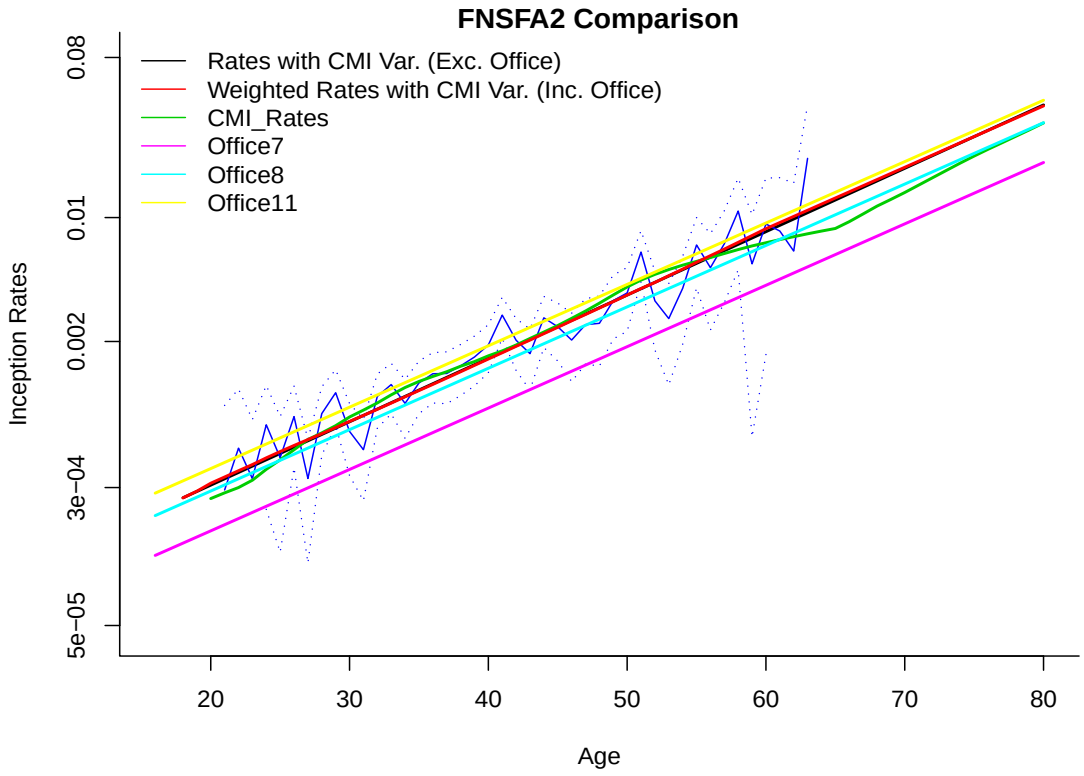


Figure 7.16: Graphs of diagnosis inception rates for females, non–smokers, full accelerated policies and durations 0 & 1.



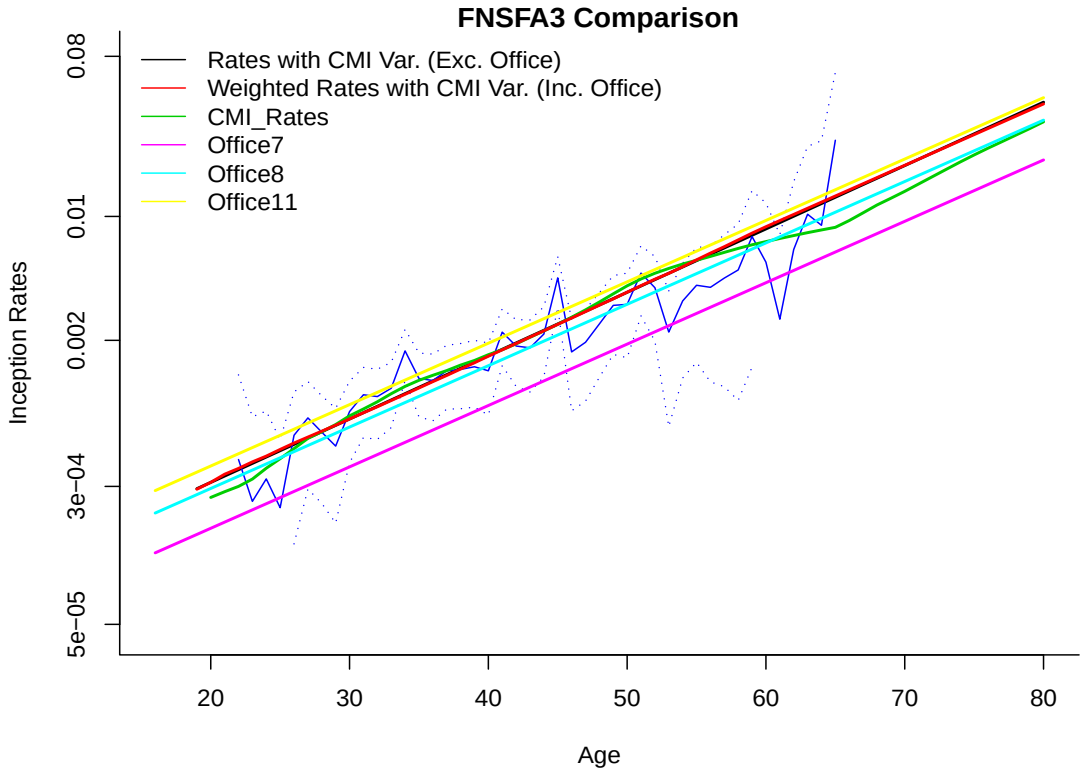


Figure 7.17: Graphs of diagnosis inception rates for females, non–smokers, full accelerated policies and durations 2 & 3.

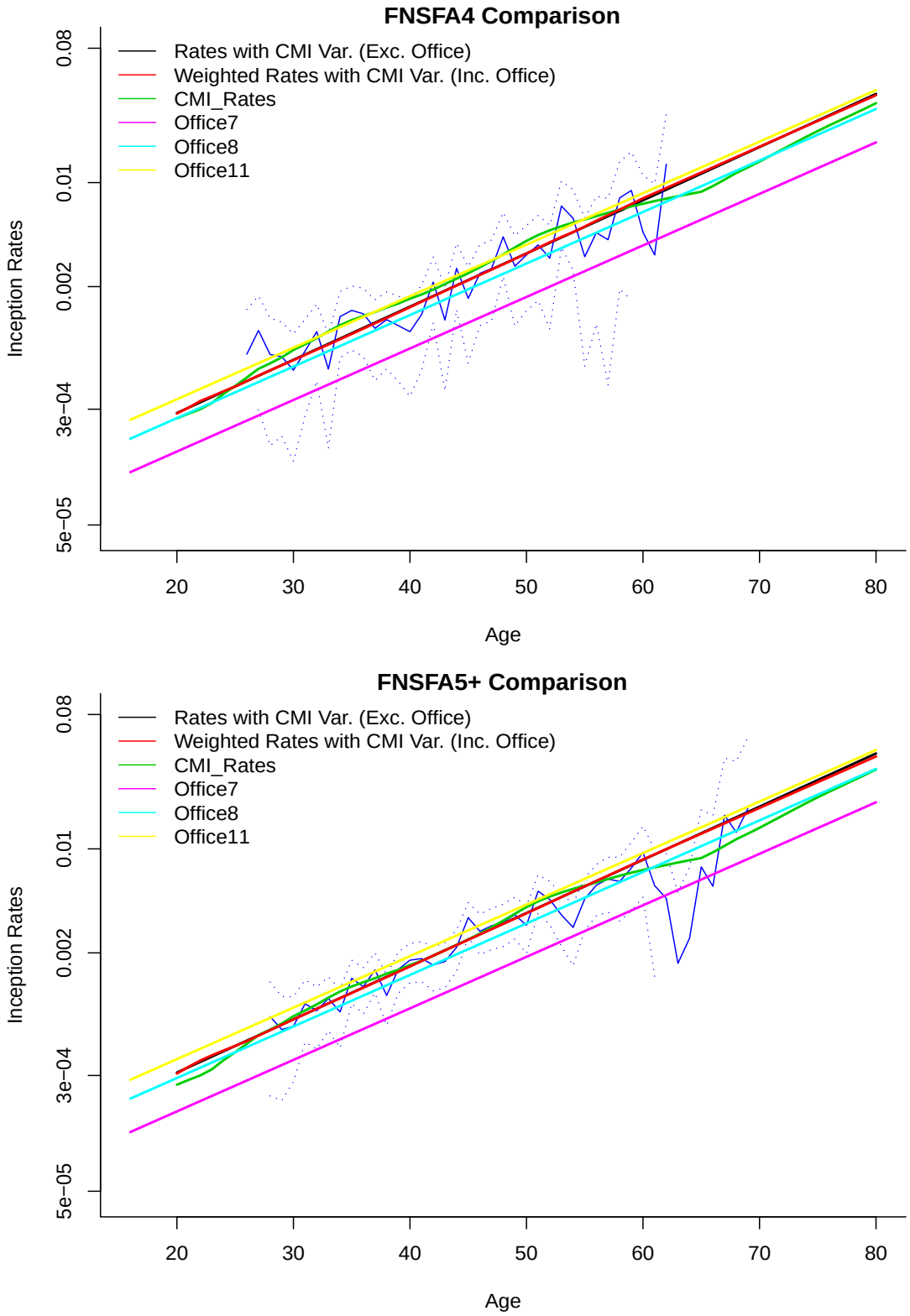


Figure 7.18: Graphs of diagnosis inception rates for females, non–smokers, full accelerated policies and durations 4 & 5+.

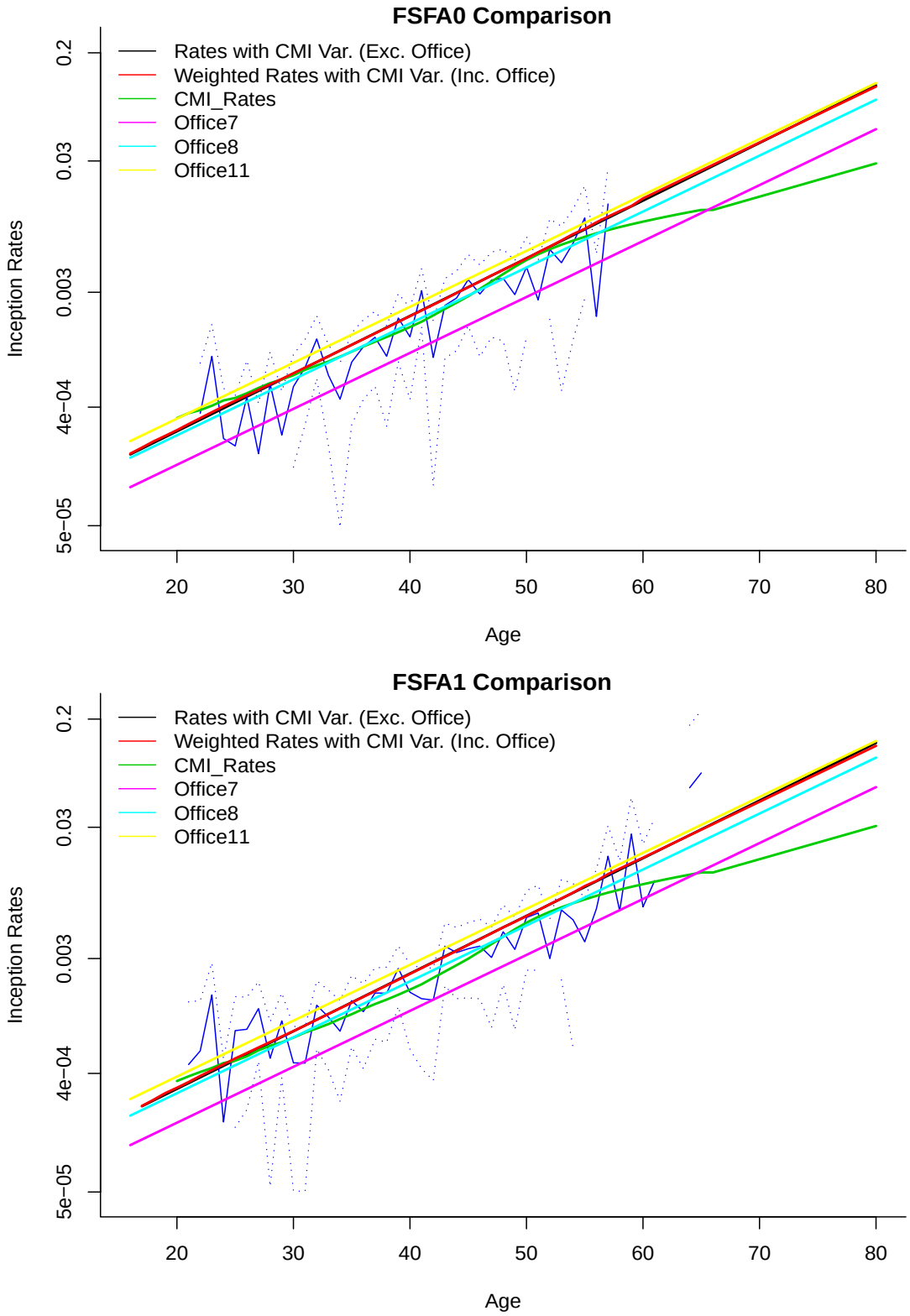


Figure 7.19: Graphs of diagnosis inception rates for females, smokers, full accelerated policies and durations 0 & 1.

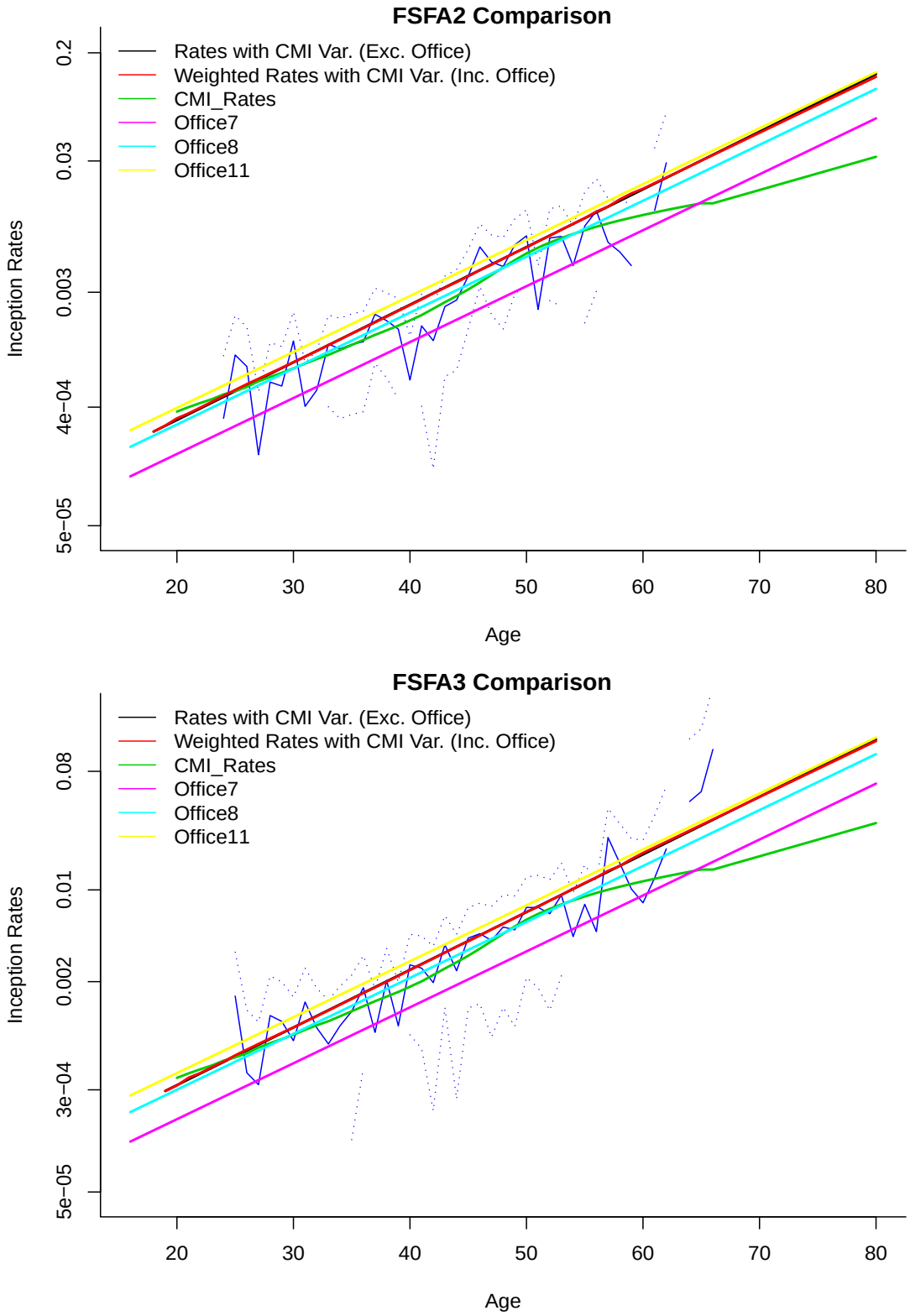


Figure 7.20: Graphs of diagnosis inception rates for females, smokers, full accelerated policies and durations 2 & 3.

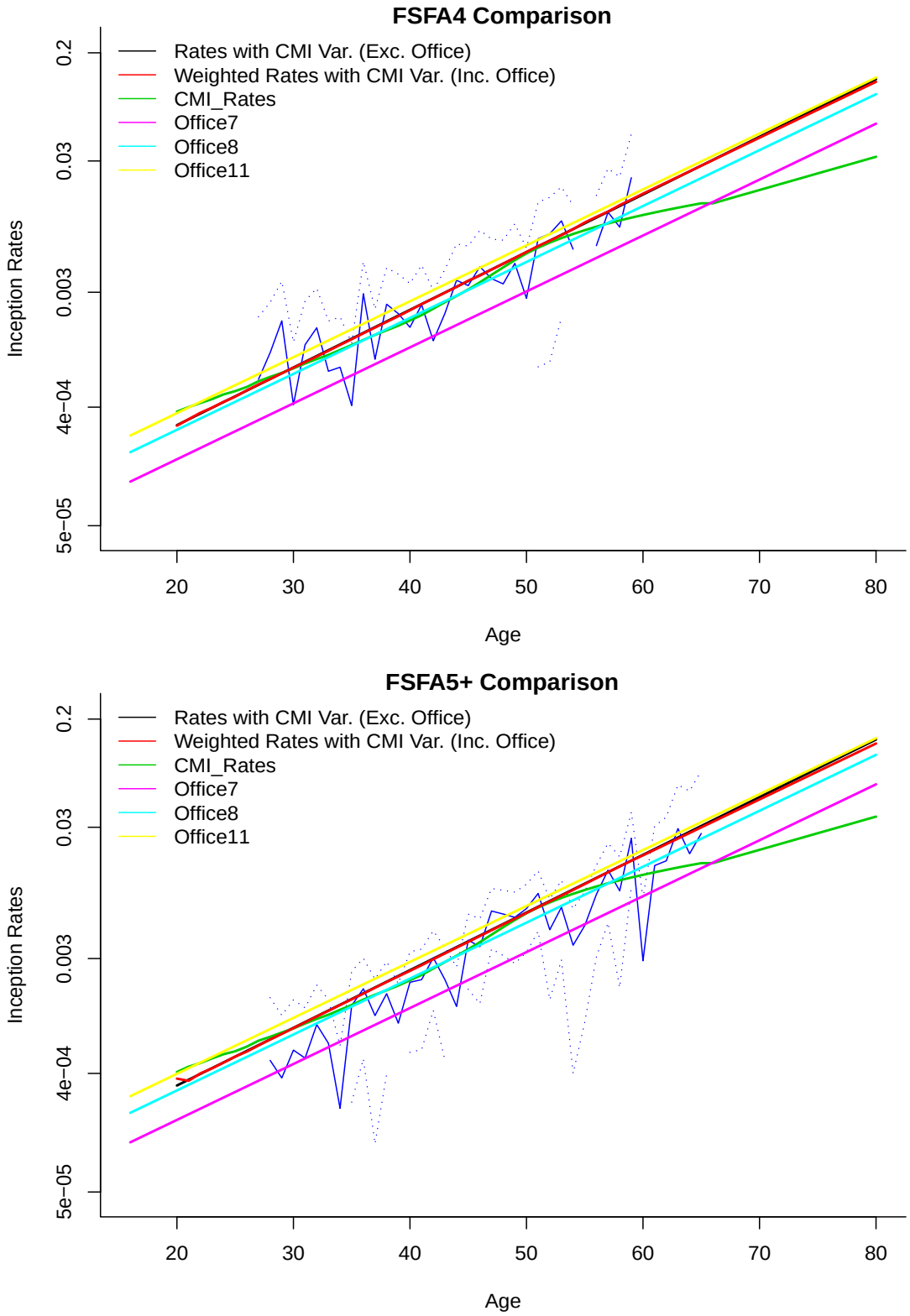


Figure 7.21: Graphs of diagnosis inception rates for females, smokers, full accelerated policies and durations 4 & 5+.

Finally, we would like to see the effect of age–smoker interaction included in the model for the inception rates. Modelled smoker rates against non–smoker rates for males, policy duration 0 are shown in Figure 7.22. As in the best model case, up to approximately age 20 the non-smoker rates are higher than the smoker rates. Since there is very little data below age 20, this ordering is unlikely to be a real feature and the rates should be adjusted for these ages. The adjustment might be increasing the inception rates for smoker to the level of inception rates for non–smokers or vice versa.

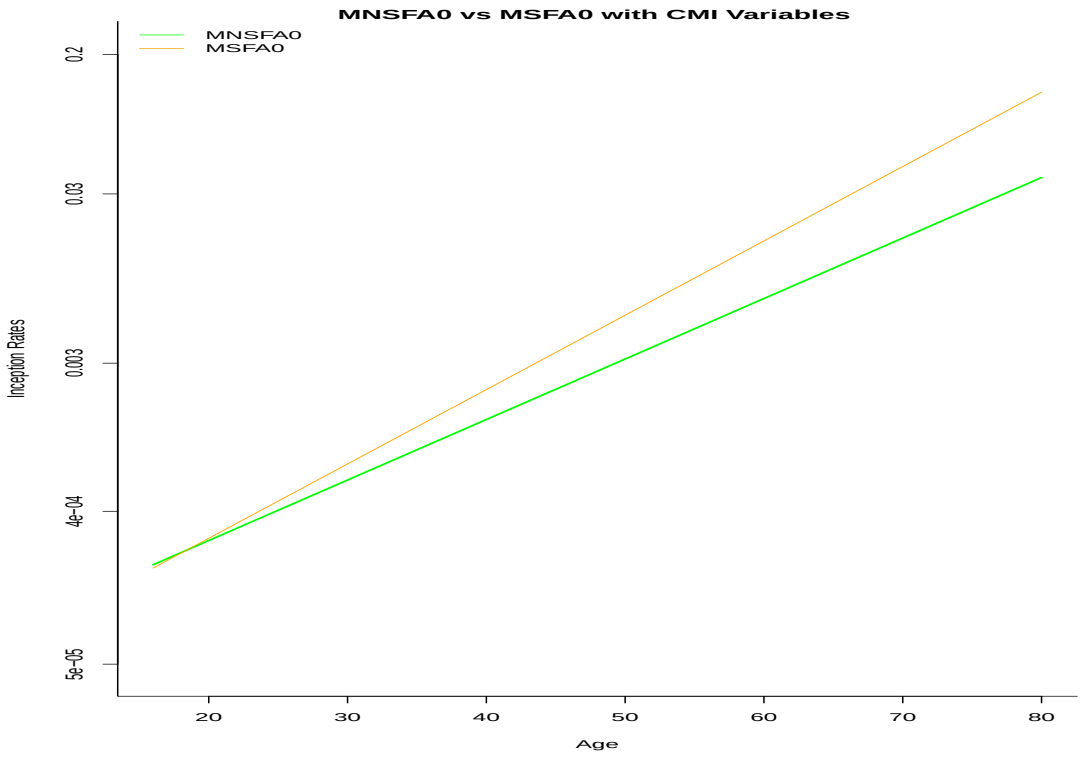


Figure 7.22: Comparison of diagnosis inception rates for non–smokers vs smokers using CMI variables for males and duration 0.

7.5 Sensitivity of the inception rates to delay estimates

In Sections 7.2 - 7.4, inception rates are calculated using the claims data for which the missing dates are estimated using the median of the CDD obtained in Section 6.3. In this section, the effect of missing date estimation on the inception rates is investigated. For different percentiles of the estimated missing delays, estimates of inception rates will be different. To show how much they change, we replace the missing delays with 2.5% and 97.5% percentiles as well as the mean of the CDD and estimate the inception rates corresponding to each of these percentiles. With the help of a confidence interval for the inception rates estimated by using the median of the CDD, we investigate whether there is a significant difference between the inception rates estimated by using the median of the CDD and the inception rates estimated using other percentiles of the CDD. The confidence interval is obtained by using a bootstrap methodology which is discussed in Section 7.5.1.

7.5.1 Sensitivity analysis

In Sections 7.2 - 7.4 we assumed the number of claim counts has a Poisson distribution with mean λE^* (see (7.1)) and we used three different models for smoothing. In this section we perform the sensitivity analysis using the best model obtained in Section 7.3. Since we use a logarithmic link function, the linear predictor of the model is in the form

$$\log \mathrm{E}(N) = \delta_{int} + \delta_{zage} x + \beta_{smoker}\theta_3 + \beta_{poldur}\theta_7 + \beta_{office}\theta_8 + \beta_{zage\times smoker} x \times \theta_3 + \log(E^*). \tag{7.9}$$

We would like to mention that the reason we use different subscripts for the $\boldsymbol{\beta}$ coefficients from the $\boldsymbol{\theta}$ covariates is convenience. Once the $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ coefficients are estimated, fitted inception rates, i.e. smoothed rates, $\tilde{\lambda}$, or fitted number of claims, $\tilde{N}$, can be calculated. In the bootstrap method we use, the number of claims can be simulated from a Poisson distribution with the fitted rates

$$N^{sim} \sim Poisson(\tilde{\lambda}E^*)$$

and we can regress N^{sim} on the covariates

$$\log(\mathrm{E}(N^{sim})) = \delta_{int}+\delta_{zage}x+\beta_{smoker}\theta_3+\beta_{poldur}\theta_7+\beta_{office}\theta_8+\beta_{zage\times smoker}x\times\theta_3+\log(E^*)$$

and repeat this to obtain a confidence interval for the inception rates (Efron and Tibshirani, 1993). We simulate 500 empirical samples with this procedure. For each of these samples, $\tilde{N}^{sim}$ and $\tilde{\lambda}^{sim}$ are calculated.

We would like to note that an alternative way of doing this is to assume that the estimated $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ coefficients in (7.9) have multivariate normal distributions (asymptotically) and to simulate coefficients from this distribution as explained by Forfar *et al.* (1988) in section 11. However, this bootstrap method uses more assumptions, such as asymptotic normality of the estimated coefficients, and it gives slightly narrower confidence intervals than the first method. We prefer to use the first method since we do not need to assume asymptotic normality.

The confidence intervals are given for the risk profiles in Section 7.3 under the best model. However, we use the age range 20 to 65 here since there are very limited or no data for some risk profiles outside of this range.

Figures 7.23 - 7.25 demonstrate the confidence intervals of the inception rates, estimated by using the median of the CDD, of the risk profiles shown previously in Figures 7.3 – 7.8 under the best model. In the figures, IR denotes the smoothed diagnosis inception rates while 2.5pc, median, 95pc and 97.5pc denote the 2.5%, 50%, 95% and 97.5% percentiles of the CDD, respectively. Missing delays are estimated using the mean of the CDD when inception rates are denoted by IR.(mean). IR.(boot.ci.median) is the confidence interval of the IR.(median) obtained by the bootstrap method. These inception rates and confidence intervals are given as a ratio of the inception rates based on the median in order to see the changes clearly. The inception rate, as estimated by using the median of the CDD, is taken to be equal to

1 i.e.

$$\text{IR.(median)} = \tilde{\lambda}/\tilde{\lambda}$$

and the other percentiles and the confidence intervals are shown as a ratio of it. For example

$$\text{IR.(mean)} = \tilde{\lambda}^{\text{mean}}/\tilde{\lambda}$$

or

$$\text{IR.(boot.ci.median)} = \tilde{\lambda}^{sim}/\tilde{\lambda}.$$

In general, the confidence intervals mostly lie within 10-15% of $\hat{\lambda}$. Especially for younger and older ages, the confidence intervals for smokers (Figures 7.26 - 7.28) are wider than for non-smokers (Figures 7.23 - 7.25).

In some of the figures we see that the rates are not very smooth for younger and older ages (see e.g. IR.(97.5pc) in Figure 7.23 or in Figure 7.27). The reason is that we use a weighted average for offices using their exposures as explained previously in this chapter and for these ages the exposure figures are not very homogeneous since we do not have enough data. This is also the reason for getting wider intervals for these ages. For example, there is a lack of smoothness in the rates at around age 20 starting from policy duration 3 (see the lower graph in Figure 7.24, graphs in Figure 7.25, the lower graph in Figure 7.27, graphs in Figure 7.28). This is because at that age a long policy duration is unusual. Therefore at around this age the confidence interval gets slightly wider and this becomes more obvious for longer policy durations such as 4 years, 5 years or more (see Figures 7.25 and 7.28). The same effect can also be seen for older policyholders with short policy durations (see Figures 7.23 and 7.26). Since we have less data for smokers, the rates are less smooth for older and younger policyholders with this risk profile.

In the upper graphs of Figures 7.23 and 7.26, which correspond to policyholders with policy duration less than a year, inception rates estimated by using the 97.5% point of the CDD lie outside of the confidence interval of IR.(median), whereas they lie inside when they are calculated using the 95% point of the CDD. The reason for that is the long tail of the Burr distribution. Note that the percentiles we use in estimation of the missing delays affect not only the missing dates of diagnosis. Since we calculate

policy duration and age at the time of diagnosis, the estimated delays also affect these variables. Because of the heavy tail of the Burr distribution, the estimated missing delays are very long using the 97.5% percentile. The average missing delay with this percentile is 645 days (with standard deviation 289.1), whereas it drops to 456 days (209.8) when the 95% point is used. Although it is still long compared to the median (115 days (83.3)) or the mean (174 days (98.6)) of the distribution, it is significantly less extreme than the 97.5% percentile. This means that, when the date of diagnosis is missing, it is estimated to be very early if the 97.5% point is used. This, in turn, means very short policy durations for these claims, where many of them fall in the less than 1 year policy duration category and it appears that most of the missing dates have policy duration less than a year.

In all graphs, the inception rates that are estimated by using the mean or 2.5% percentile are very close to those estimated by using the median of the CDD. For longer policy durations, the inception rates based on the 97.5% percentile of the CDD are closer to the other inception rates estimated using other CDD percentiles.

From the sensitivity analysis in this section, we can conclude that the inception rates are not very sensitive to the point estimate used to obtain the missing dates. Since using a reasonable percentile of the CDD (between 2.5% and 95%) for the missing dates will give an inception rate within the 95% confidence interval of IR.(median), the median of the CDD can be safely used to estimate the missing dates.

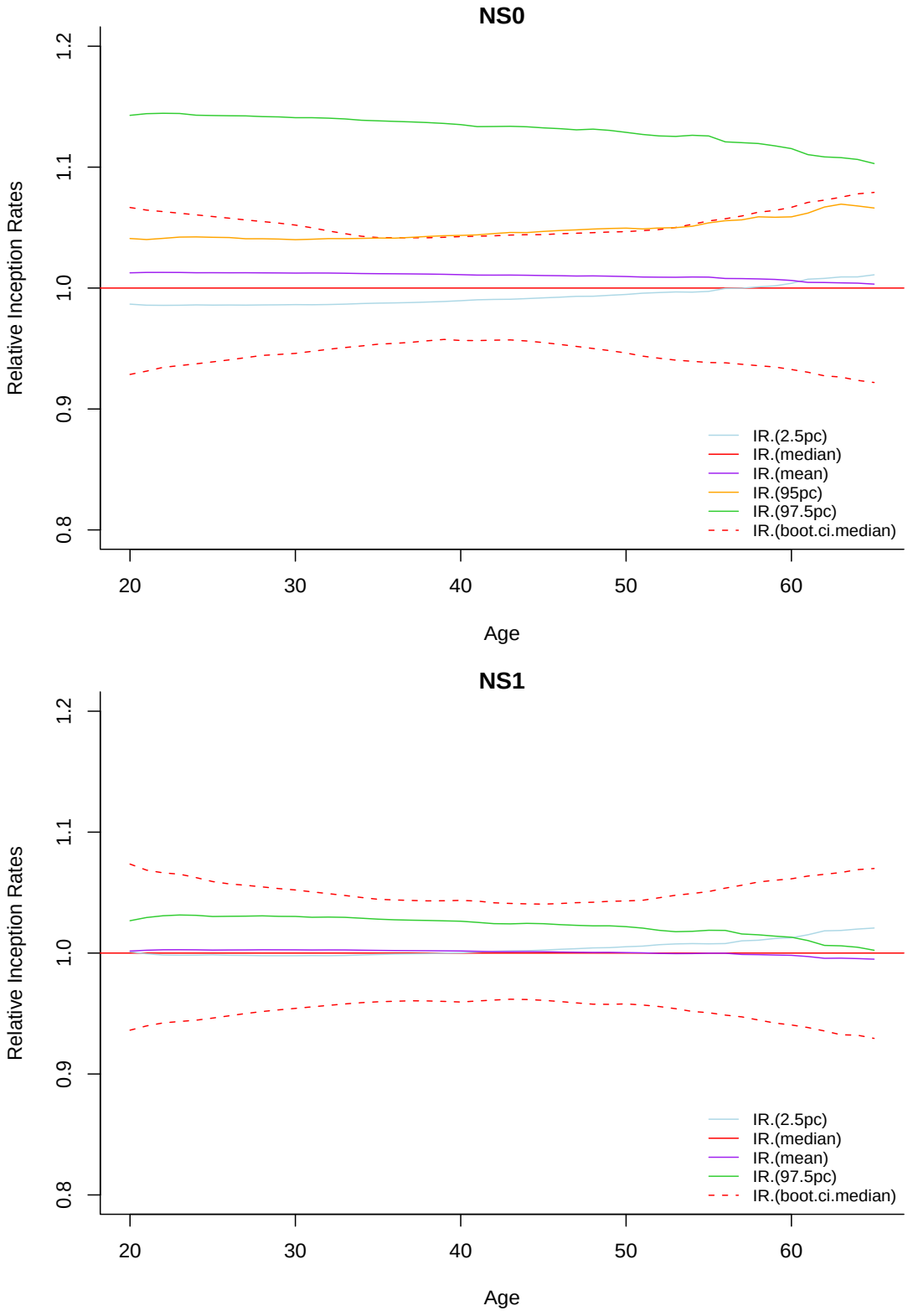


Figure 7.23: Graphs of relative diagnosis inception rates with different missing delay estimates and confidence intervals of the inception rates based on the median of the CDD (as a ratio of inception rates based on the median of the CDD) for non–smokers and durations 0 & 1.

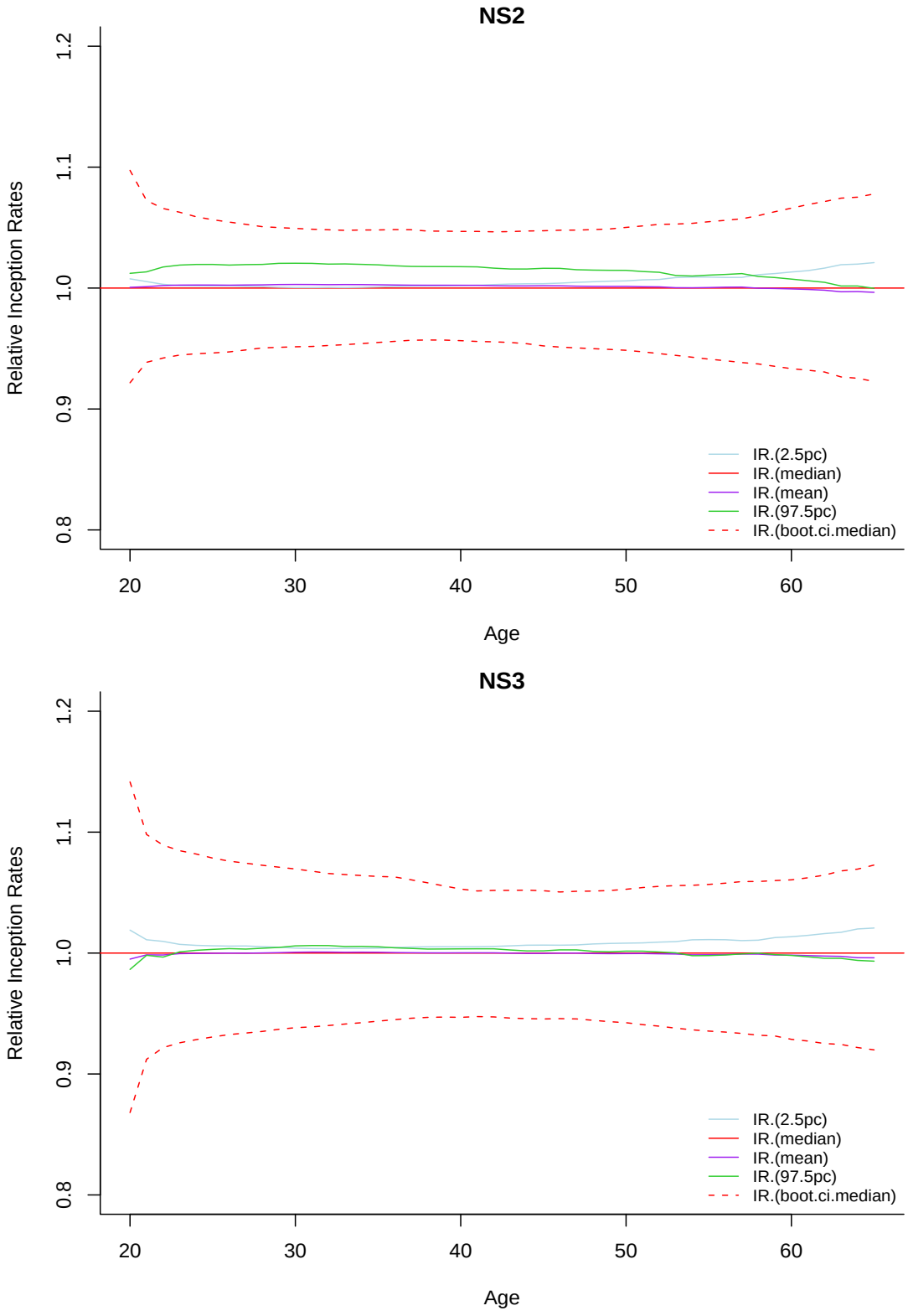


Figure 7.24: Graphs of relative diagnosis inception rates with different missing delay estimates and confidence intervals of the inception rates based on the median of the CDD (as a ratio of inception rates based on the median of the CDD) for non–smokers and durations 2 & 3.

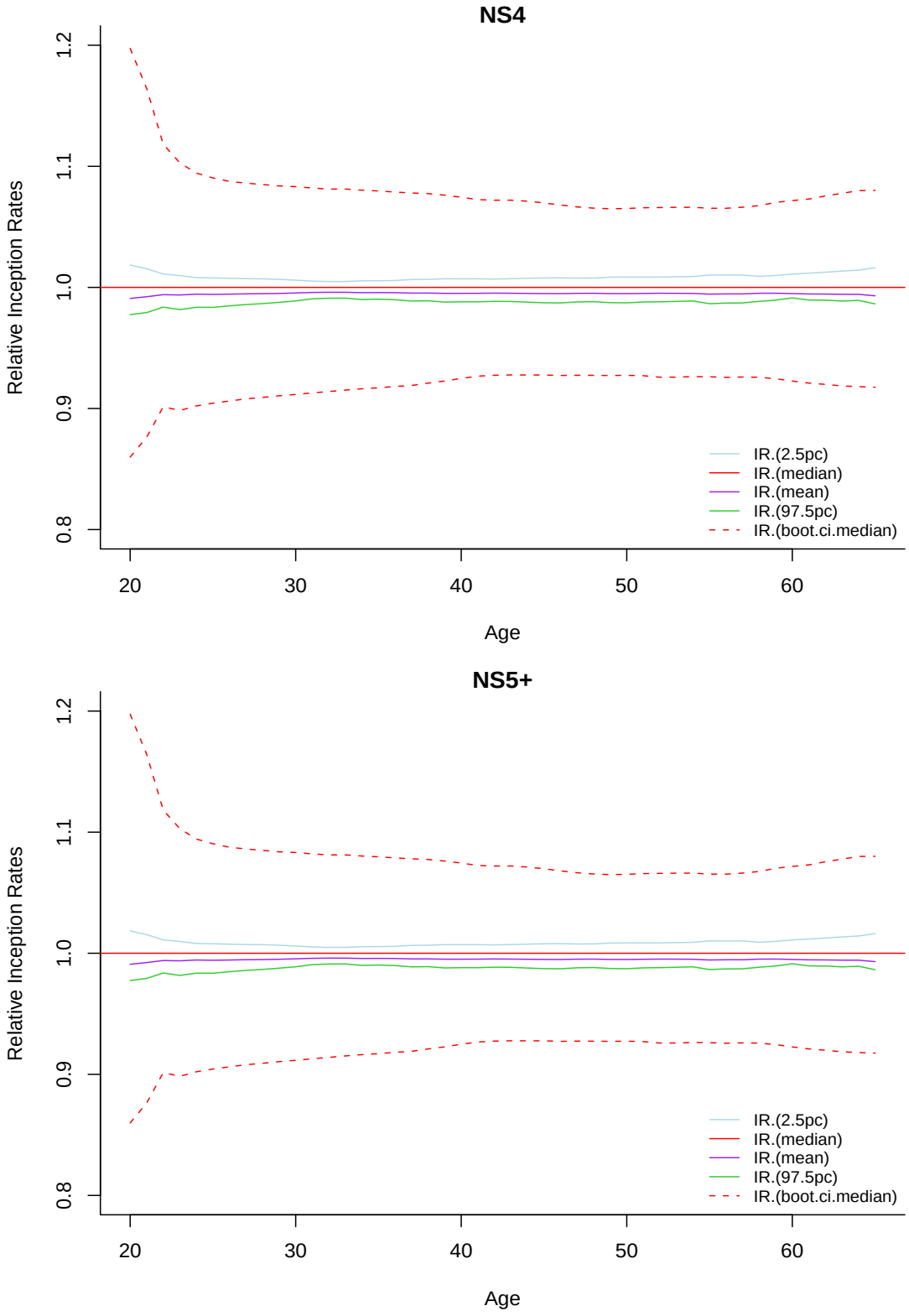


Figure 7.25: Graphs of relative diagnosis inception rates with different missing delay estimates and confidence intervals of the inception rates based on the median of the CDD (as a ratio of inception rates based on the median of the CDD) for non–smokers and durations 4 & 5+.

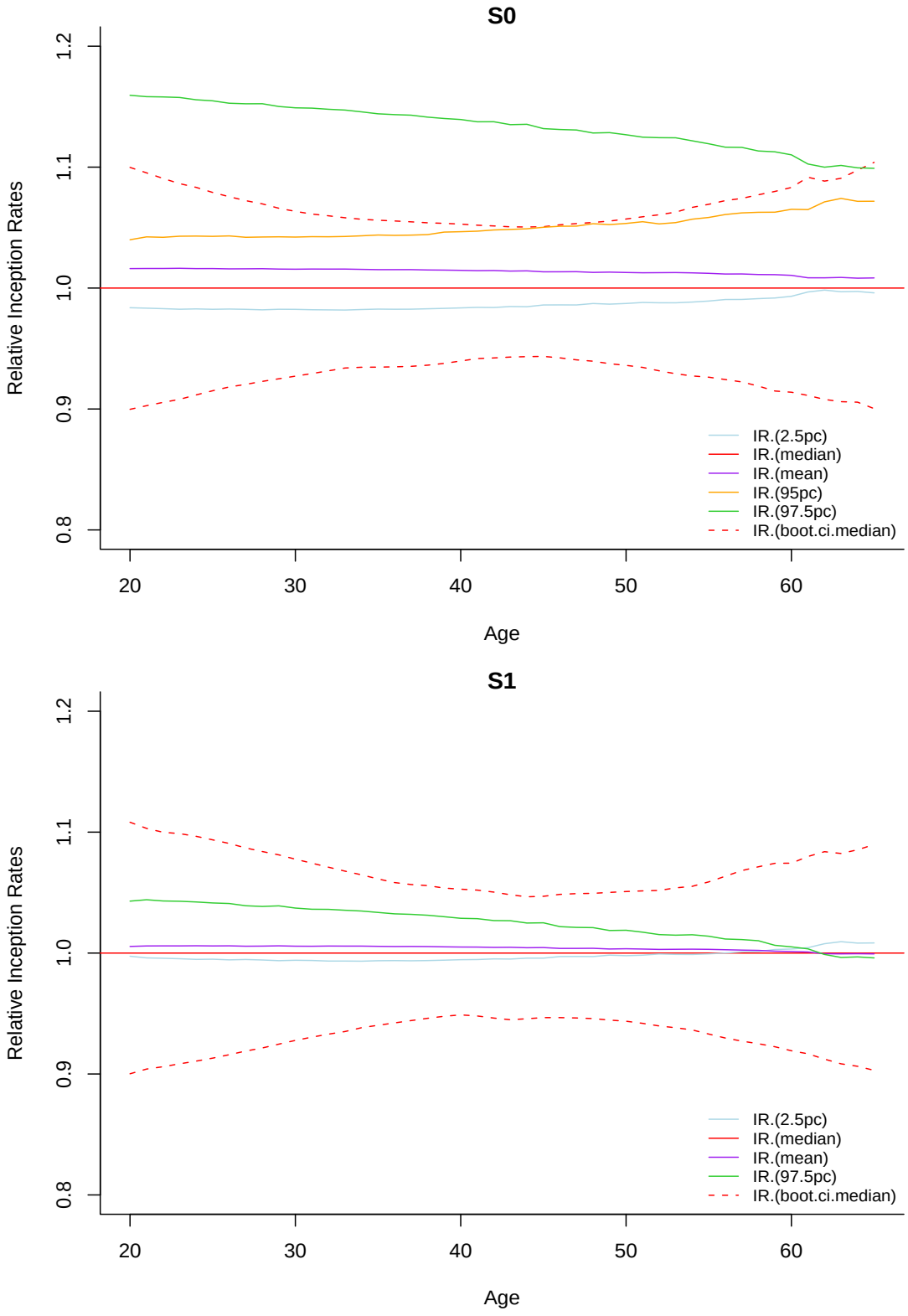


Figure 7.26: Graphs of relative diagnosis inception rates with different missing delay estimates and confidence intervals of the inception rates based on the median of the CDD (as a ratio of inception rates based on the median of the CDD) for smokers and durations 0 & 1.

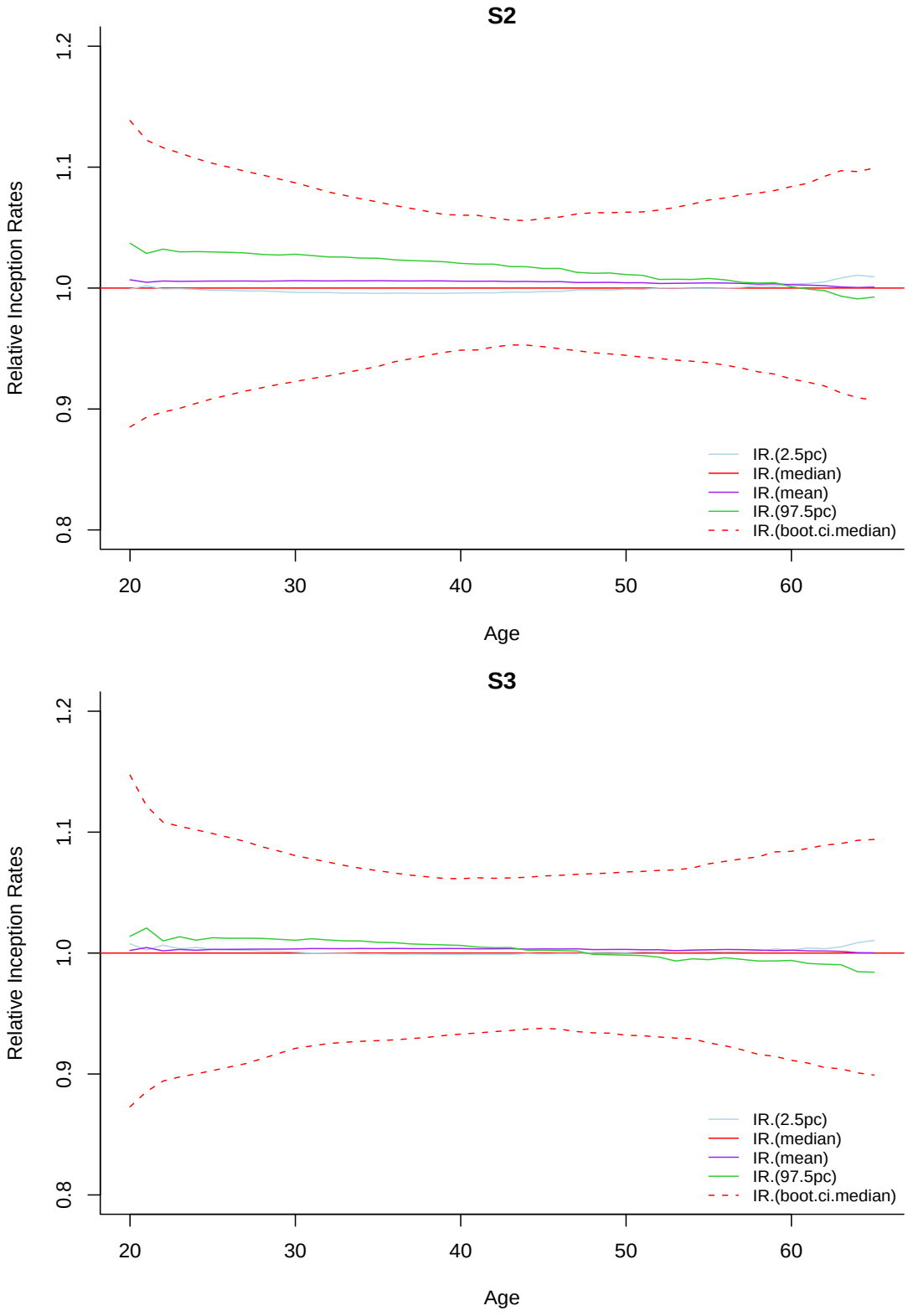


Figure 7.27: Graphs of relative diagnosis inception rates with different missing delay estimates and confidence intervals of the inception rates based on the median of the CDD (as a ratio of inception rates based on the median of the CDD) for smokers and durations 2 & 3.

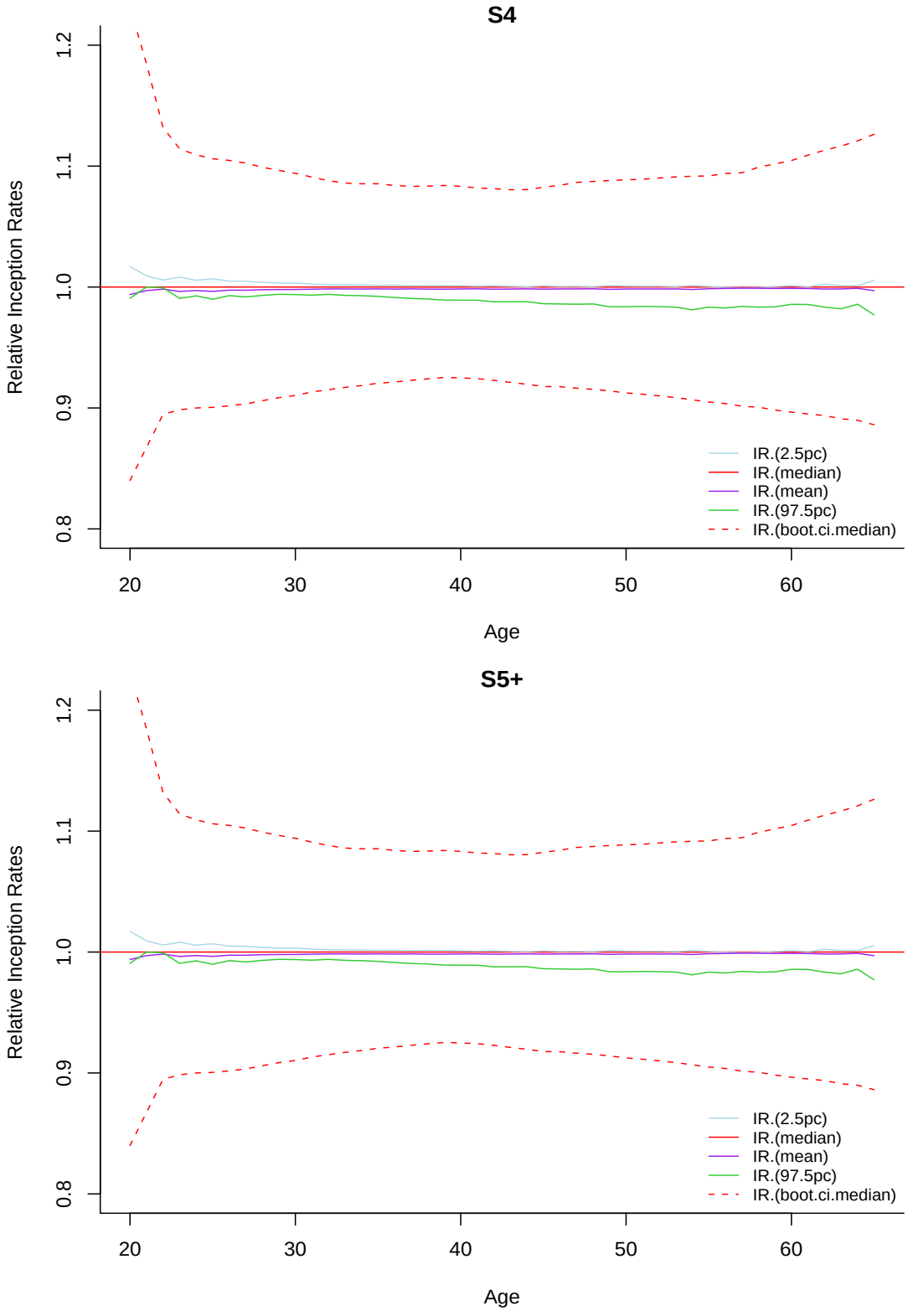


Figure 7.28: Graphs of relative diagnosis inception rates with different missing delay estimates and confidence intervals of the inception rates based on the median of the CDD (as a ratio of inception rates based on the median of the CDD) for smokers and durations 4 & 5+.

Chapter 8

Diagnosis inception rates II: Cause-specific rates

In this chapter we derive the intensity rates of a diagnosis of a sickness at age x last birthday with characteristics $\boldsymbol{\theta}$ which will lead to a claim from particular cause c. This intensity is denoted by $\lambda^c(x, \boldsymbol{\theta})$. The same characteristics vector, $\boldsymbol{\theta}$, given in Table 7.1 is used. Again, we restrict the age to $16 - 80$ for the analyses. Also, the effect of the age - smoker interaction is investigated for the same reason stated in Section 7.3. The rates are provided for the main claim causes, i.e. CABG, cancer, death, heart attack, kidney failure, major organ transplantation, MS, Stroke and TPD, and for other causes. Variable selection is performed in the same way as described in Section 7.2. Note that we start model selection with $g_0(x), f_2(x)$ function. Here, we do not force the first order exponential age term to stay in the model but it stays in models for almost all causes as age is an important factor for the morbidity rates. On the other hand, if it drops from the model (e.g. MOT, MS), we select the best model between models with $g_0(x), f_1(x)$ and $g_1(x), f_1(x)$ functions.

For some causes for male-non–smokers and different policy durations we show the CMI rates presented in WP 43 (2010) for comparison purposes. Although the CMI rates are provided from age 20 to 80, they mention that for the specific causes the rates outside the age range 31 and 60 are only indicative.

Goodness of fit of models is assessed using Pearson's χ^2. For almost all causes, p-values were found above 1%. Two exceptions are cancer (p-value $= 4.88 \times 10^{-6}$) and

death (p-value = 1.15×10^{-11}). We note that goodness of fit is not easily achieved when the aim is to obtain smoothed diagnosis inception rates, especially in large data sets.

The structure of the model is explained in Section 8.1 and in Section 8.2 the best models for individual causes are presented. In Section 8.3 these individual causes are compared to the all-cause rates obtained in Chapter 7.

8.1 Structure of the model

We define $N^c(x, \boldsymbol{\theta})$ as the observed number of claims for cause c at age x and risk profile $\boldsymbol{\theta}$ which are diagnosed in year θ_5 and settled before the end of the last contribution year of the considered office, t_{θ_8}, and we assume a Poisson distribution so that

$$N^c(x, \boldsymbol{\theta}) \sim Poisson(\lambda^c(x, \boldsymbol{\theta}) E^{*c}(x, u; \boldsymbol{\theta})) \tag{8.1}$$

where $\lambda^c(x, \boldsymbol{\theta})$ denotes the intensity for a claim diagnosis in the respective year θ_5 from cause c at age x last birthday and risk profile $\boldsymbol{\theta}$, and the adjusted exposure $E^{*c}(x, u; \boldsymbol{\theta})$ is

$$E^{*c}(x, u; \boldsymbol{\theta})) = \int_{u=0}^{1} E(x, u; \boldsymbol{\theta})\, F^c(t_{\theta_8} - \theta_5 + 1 - u; x, \boldsymbol{\theta})\, du. \tag{8.2}$$

In (8.2), the adjustment factor, $F^c(t_{\theta_8} - \theta_5 + 1 - u; x, \boldsymbol{\theta})$, denotes the probability of settling a claim in time $(t_{\theta_8} - \theta_5 + 1 - u)$, at age x last birthday, given risk profile $\boldsymbol{\theta}$ and cause c by the end of year t_{θ_8} starting from time u after the start of year θ_5, where $\theta_5 = 1999, \ldots, 2004, 2005$ and $t_{\theta_8} \in (\theta_5, \theta_5 + 1, \ldots, 2005)$ and the exposure, $E(x, u; \boldsymbol{\theta})$, is calculated in the same way as explained in Section 7.2. The integral is approximated for each cause seperately by using a four-step Simpson's Rule.

Under the Poisson distribution given in (8.1), the crude estimator for the intensity and the standard error of the estimate is given in (8.3) and (8.4), respectively.

$$\hat{\lambda}^{raw,c}(x,\boldsymbol{\theta}) = N^c(x,\boldsymbol{\theta}) \Big/ E^{*c}(x,u;\boldsymbol{\theta})\,du \tag{8.3}$$

$$sd(\hat{\lambda}^{raw,c}(x,\boldsymbol{\theta})) = \sqrt{N^c(x,\boldsymbol{\theta})} \Big/ \int_{u=0}^{1} E^{*c}(x,u;\boldsymbol{\theta})\,du. \tag{8.4}$$

For smoothing the crude rates we use the same function given in (7.7) with Poisson errors.

8.2 Best models for specific causes

Coronary artery bypass graft (CABG)

The two risk factors for coronary heart diseases that one can not change are aging and being male. Smoking is also known as a serious risk factor for heart diseases (Chatterjee *et al.*, 2009). Under different age polynomials for CABG, the gender and smoking status covariates are always found to be important after model selection using stepwise regression where BIC is used as a selection criterion. The models we have considered can be seen in Figure 8.1. This figure also shows the path we used in model selection.

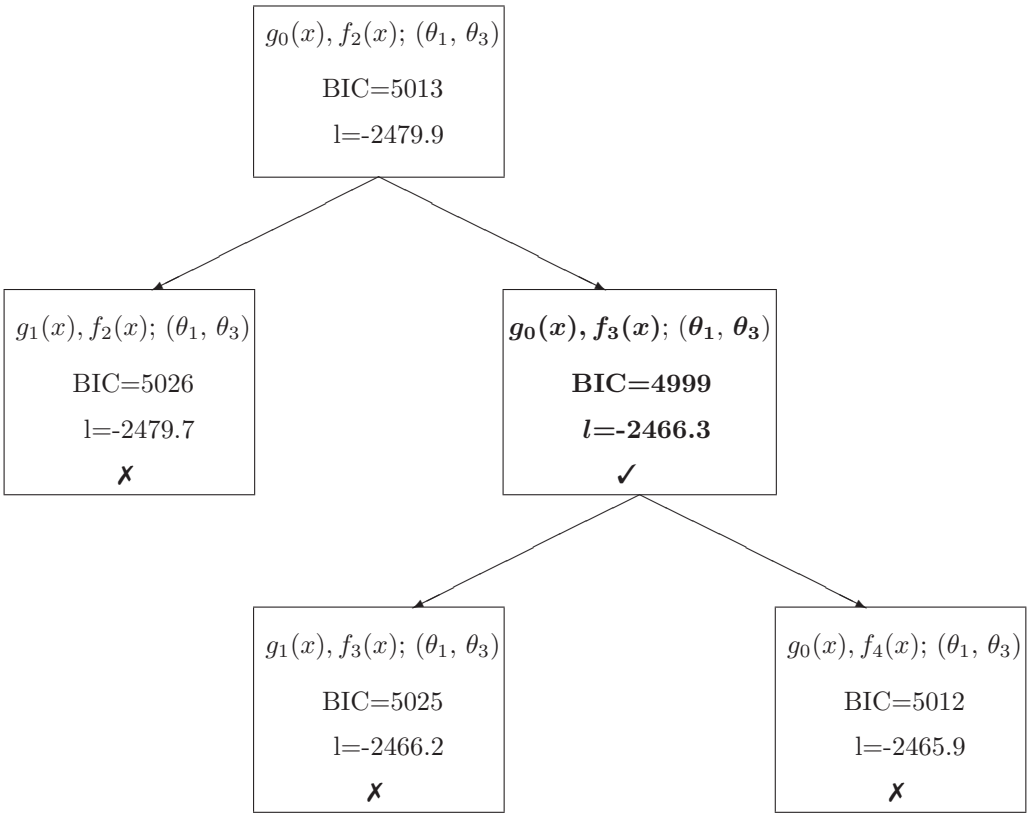


Figure 8.1: Model selection for CABG.

The smallest BIC is obtained under the quadratic exponential polynomial for age $(g_0(x), f_3(x))$ with the sex (θ_1) and smoking status (θ_3) covariates. The function given in 8.5 is used to smooth the CABG rates

$$\lambda^{CABG} = \exp\left(\delta_{int} + \delta_{zage}x + \delta_{zage^2}x^2 + \beta_{sex}\theta_1 + \beta_{smoker}\theta_3\right) \tag{8.5}$$

and parameter estimates are given in Table 8.1. As expected, males and smokers have higher CABG rates compared to the base categories of females and non–smokers. Age has a positive effect on the intensity rates, however the negative coefficient of age^2 indicates a decreasing age effect for older ages (Note here that age is standardised by subtracting the mean (39.75) and dividing by the standard deviation (11.21) as stated in Chapter 7).

Table 8.1: ML estimates of parameters under the best model for CABG.

Parameter	Estimate	Std. Error	p-value
$\delta_{intercept}$	-12.8404	0.2148	$< 2 \times (10^{-16})$
δ_{zage}	5.6621	0.8185	$4.6 \times (10^{-12})$
δ_{zage^2}	-3.1830	0.6858	$3.5 \times (10^{-6})$
β_{sex}	1.7691	0.1696	$< 2 \times (10^{-16})$
β_{smoker}	0.5814	0.1174	$7.4 \times (10^{-7})$

Figures 8.2 and 8.3 show the smooth and crude rates for CABG. Note that, in this chapter, the rates are provided in the log scale in figures, in order to be able to see the details, especially for younger ages and the y-axis is presented in original scale. Therefore, intervals between labels are not equally spaced.

In the graphs it can be seen that rates are almost flat after age 65. Like all surgical operations, CABG has a risk of complications. Some serious risks such as stroke and heart attack are more likely for elderly people (NHS, 2010). Also, comorbidity is more common for older people which increases the chance of developing complications. Therefore, in the literature, it is controversial to operate on elderly people (see e.g. MacDonald *et al.* (2000), Bowling *et al.* (2001), Bradshaw *et al.* (2001)). It appears that more cautious approaches to treatments are considered for these people, such as medical treatments rather than surgical ones (Stone *et al.*, 1996; Bowling *et al.*, 2001), and this might be one of the reasons for lower surgical rates at older ages. In Figure 8.2, for non–smokers, the CMI rates in WP 43 (2010) are also provided. Decreasing CABG rates for older ages can be seen from these rates as well.

We also mention that there are very few CABG claims for females as can be seen from Figure 8.3. Therefore none of the lower bounds of the confidence interval could be provided.

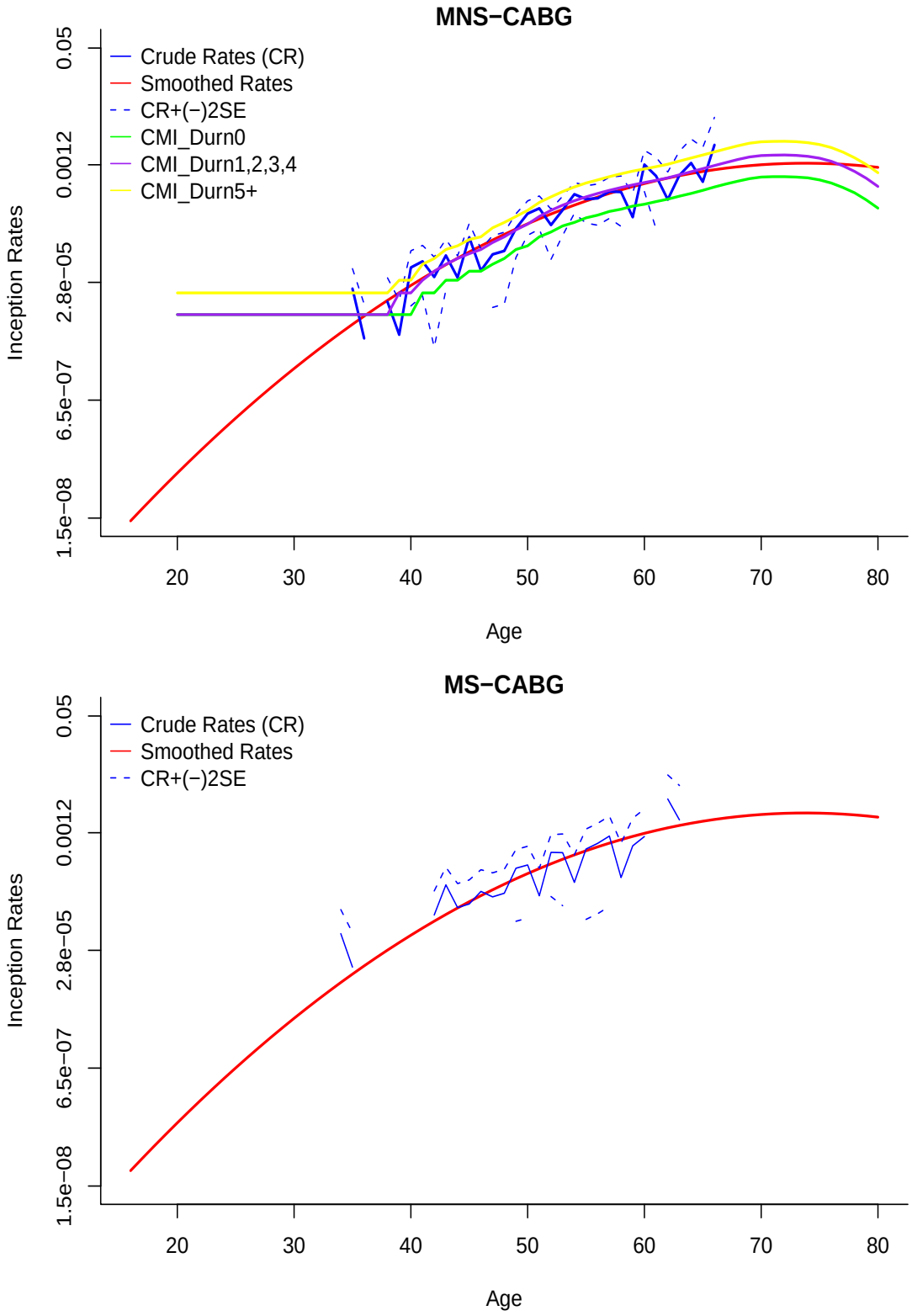


Figure 8.2: Graphs of diagnosis inception rates for CABG for males, non–smokers (MNS) and smokers (MS).

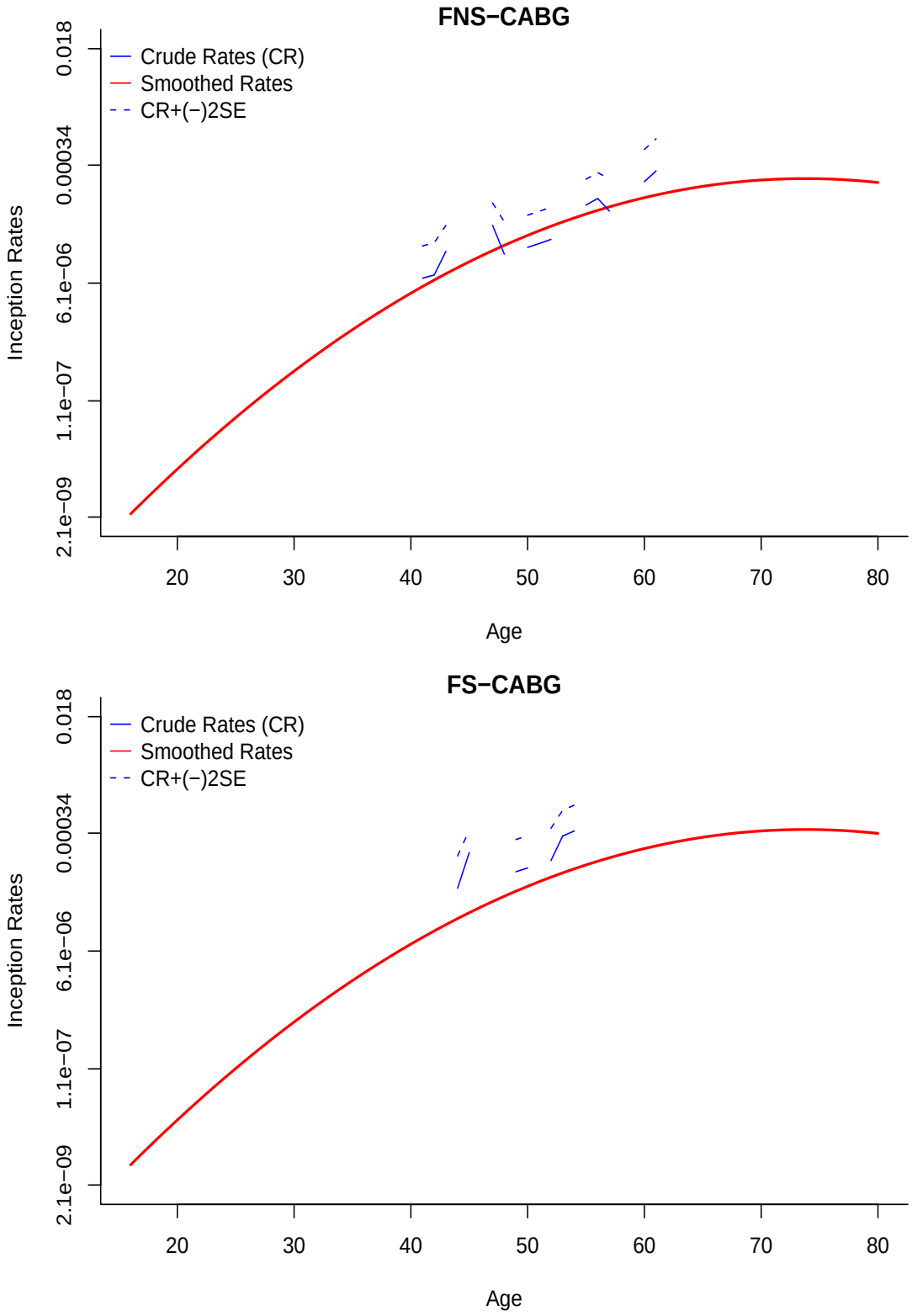


Figure 8.3: Graphs of diagnosis inception rates for CABG for females, non–smokers (FNS) and smokers (FS).

Cancer

As stated before, cancer claims form almost half (49%) of the claims data. This is not surprising, considering the statistics given by Cancer Research UK (2010) which states that more than one in three people in the UK will suffer from cancer during their lives.

After model selection under different age polynomials, sex, smoker status and year covariates have remained in the models. The age - smoker interaction term, found to be important for the model with $g_0(x), f_2(x)$ age function, dropped from the model with $g_0(x), f_2(x)$ age function. Figure 8.4 summarises the models we have considered for smoothing the cancer rates.

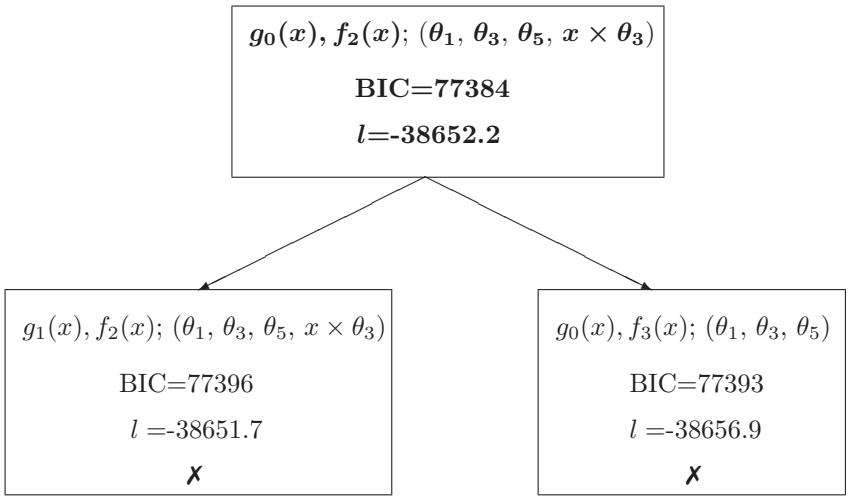


Figure 8.4: Model selection for cancer.

We started with a model with a first order exponential age function $(g_0(x), f_2(x))$. Adding a linear intercept term to this model $(g_1(x), f_2(x))$, or the model obtained after increasing the exponential order of age by one $(g_0(x), f_3(x))$ did not give a better BIC value. We use the model given in (8.6) for smoothing the cancer rates. The estimates of the model parameters are presented in Table 8.2. The effect of age and smoking are found to be positive. The intensity for males is lower than females, probably

because of high breast cancer rates for females. Finally the effect of diagnosis year is found to be important and positive, implying higher cancer rates for more recent years (note that year is standardised by subtracting the mean (2002.36) and dividing by the standard deviation (1.86)). This fact is also noted by Cancer Research UK (2010). According to that study, it is mentioned that although the rates have not changed very much within the last decade, cancer rates in the UK increased by 25% between 1978 and 2007. This rise is 14% for men and 32% for women. The effect of age - smoker interaction will be shown in Figure 8.8 later.

$$\lambda^{Cancer} = \exp\left(\delta_{int} + \delta_{zage}x + \beta_{sex}\theta_1 + \beta_{smoker}\theta_3 + \beta_{zyear}\theta_5 + \beta_{zage\times smoker}x \times \theta_3\right) \tag{8.6}$$

Table 8.2: ML estimates of parameters under the best model for cancer.

Parameter	Estimate	Std. Error	p-value
$\delta_{intercept}$	-6.8689	0.0153	$< 2 \times (10^{-16})$
δ_{zage}	0.9357	0.0143	$< 2 \times (10^{-16})$
β_{sex}	-0.5595	0.0217	$< 2 \times (10^{-16})$
β_{smoker}	0.0703	0.0287	0.0142
β_{zyear}	0.0514	0.0119	$< 1.5 \times (10^{-5})$
$\beta_{zage\times smoker}$	0.1201	0.0326	0.0002

Figure 8.5 shows the crude and modelled cancer rates for males, non-smokers. These rates are the weighted average over year. Since year is an important covariate in the model, cancer rates depend on it. However, we would like to produce cancer rates independent of the year. Therefore, we use a weighted average where the rates are weighted by exposure (similarly to what we have done in Chapter 7 to find weighted rates for offices, see (7.8)).

To find weighted smoothed rates for year (red lines in Figures 8.5 - 8.7), after obtaining the year-specific inception rates from the model given in (8.6) we weight these rates with year-specific exposures with the same characteristics and divide by the sum of these exposures. This can be expressed as

$$\frac{\sum_{i=1999}^{2005} \hat{\lambda}(x;\theta_{5,i};\theta_{\setminus\theta_5})E^*(x;\theta_{5,i};\theta_{\setminus\theta_5})}{\sum_{i=1}^{7} E^*(x;\theta_{5,i};\theta_{\setminus\theta_5})} \tag{8.7}$$

where $\theta_{5,i}$ denotes year_i for $i = 1999 \ldots 2005$ and $\theta_{\setminus\theta_5}$ denote the remaining characteristics found important for cancer except year. We have used the same procedure to find the weighted crude rates.

In Figure 8.5, the weighted crude and smoothed rates for males - non–smokers are presented together with CMI rates for different policy durations. Figures 8.6 and 8.7 show the smooth and crude rates for cancer for other combinations of gender and smoking status. In these figures we also show the smoothed rates for individual year effects for years 1999 (lowest effect) and 2005 (highest effect) to see the effect of the years. The weighted rates are closer to year 2005 compared to 1999 since we have more data for the recent years.

The confidence intervals are wider for ages below 30 and above 60 as we have less data compared to the age range 30–60. Also, since we have more data for non–smokers the confidence intervals (upper graphs of Figures 8.6 and 8.7) are tighter than the confidence intervals of smokers (lower graphs of Figures 8.6 and 8.7).

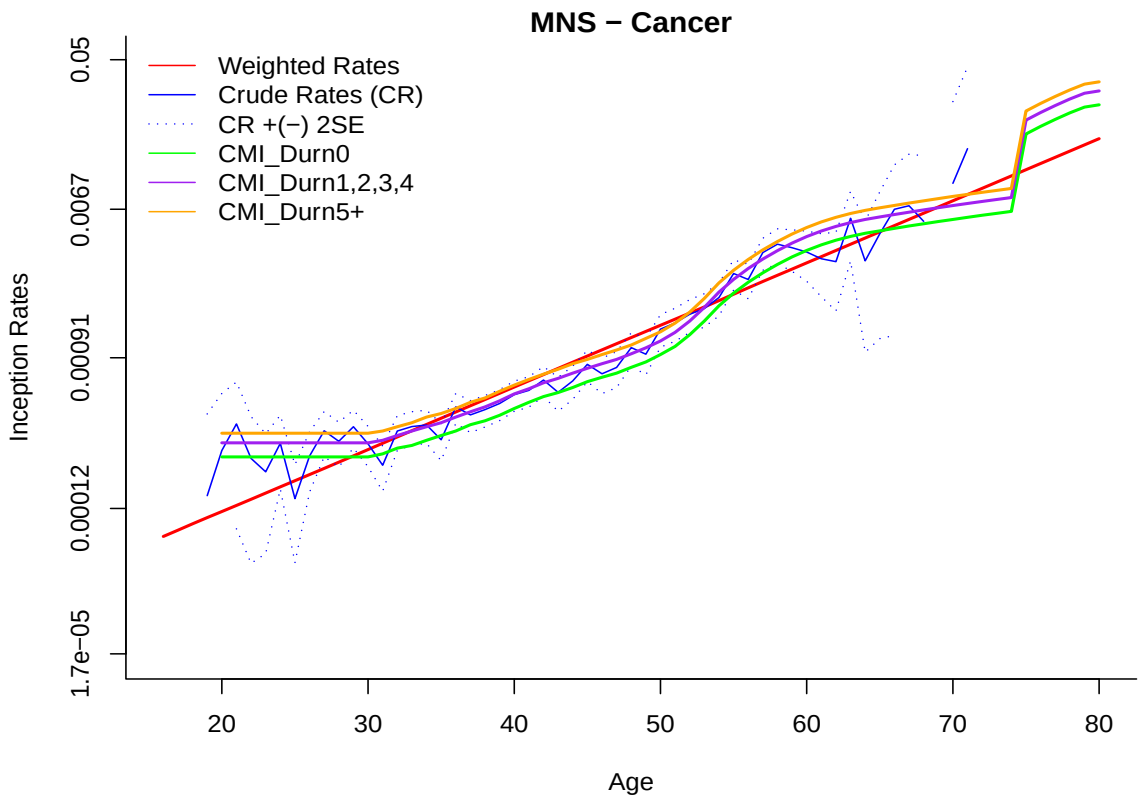


Figure 8.5: Graphs of diagnosis inception rates for cancer for males, non–smokers (MNS) together with CMI rates.

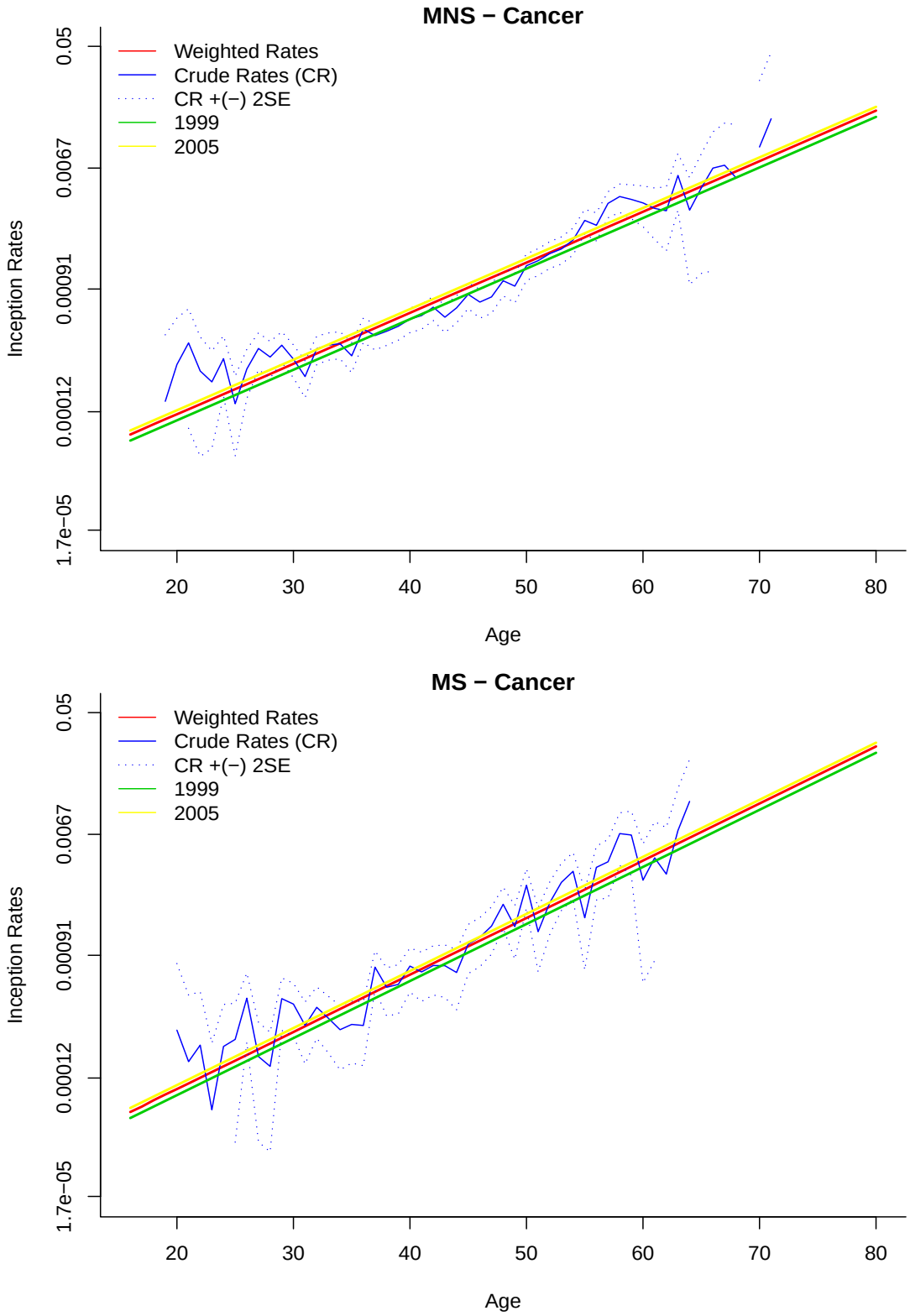


Figure 8.6: Graphs of diagnosis inception rates for cancer for males, non–smokers (MNS) and smokers (MS).

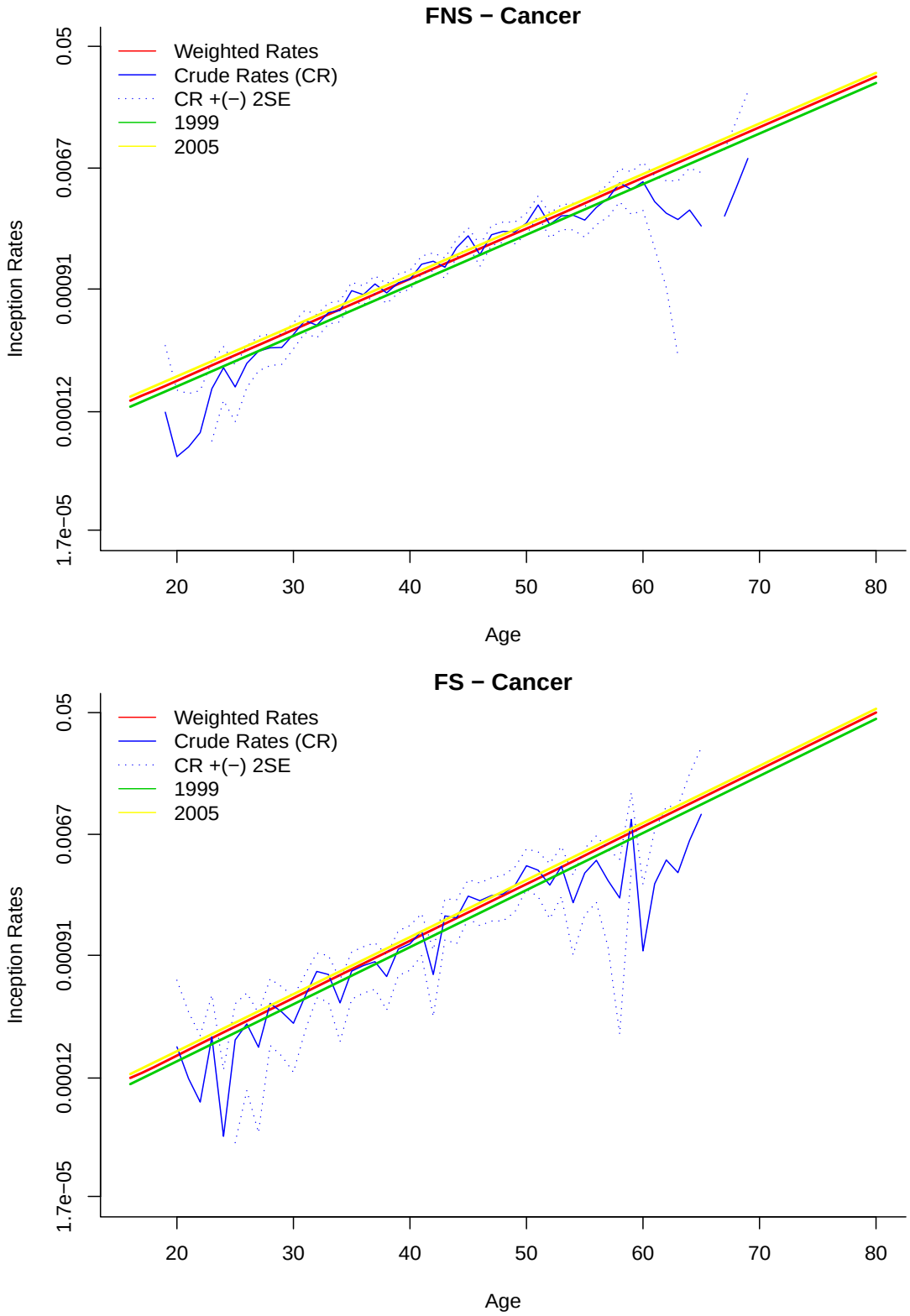


Figure 8.7: Graphs of diagnosis inception rates for cancer for females, non–smokers (FNS) and smokers (FS).

Modelled smoker rates against non–smoker rates for males are shown in Figure 8.8 to see the effect of age - smoker interaction. From about age 32 the incidence rates for smokers are higher than for non–smokers, as expected. Thus below age 30 the model fit is poor as we do not expect smoker rates to be lower than non–smoker rates. One possible reason for this might be lack of data for this age range. In general, we do not have much data for younger ages and this is more evident for individual causes.

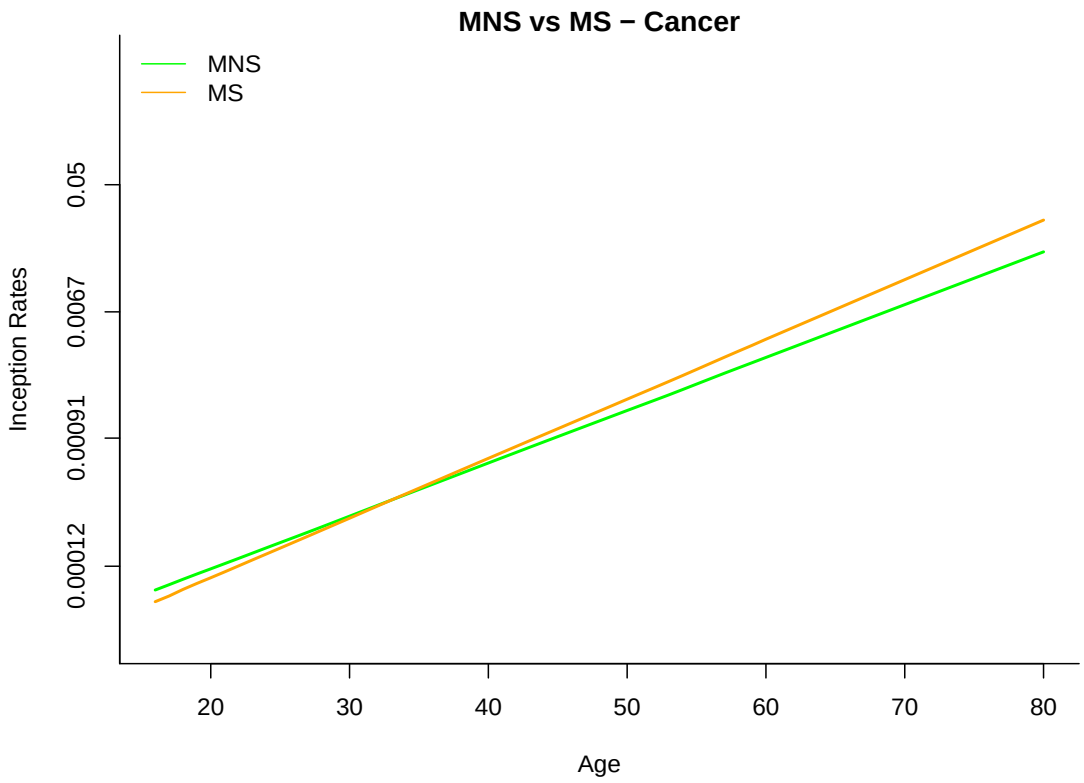


Figure 8.8: Graphs of diagnosis inception rates for cancer for males, non–smokers vs smokers.

Death

Since the death claims are covered only in fully accelerated policies, we work with a subset of data which has 359 587 different combinations of characteristics. Sex (θ_1), smoker (θ_3) and age - smoker interaction ($x \times \theta_3$) are found to be the most important covariates in modelling the mortality rates for each of the different age polynomials. Unlike all other causes, introducing a linear constant to the model improves the fit significantly. This age-independent linear constant corresponds to the Makeham constant and captures external causes of deaths such as accidents. The model selection procedure is described in Figure 8.9.

The model with $(g_1(x), f_3(x))$ age polynomials has the smallest BIC value and is in the form

$$\lambda^{Death} = \kappa_{int} + \exp\left(\delta_{int} + \delta_{zage}x + \delta_{zage^2}x^2 + \beta_{sex}\theta_1 + \beta_{smoker}\theta_3 + \beta_{zage\times smoker}x \times \theta_3\right). \tag{8.8}$$

The estimates of the model parameters are given in Table 8.3. According to this model, males have higher mortality rates than females. The death rates for smokers are increasing by age. The mortality rates increase considerably at older ages.

Table 8.3: ML estimates of parameters under the best model for death.

Parameter	Estimate	Std. Error	p-value
$\kappa_{intercept}$	-0.0002	$2.4 \times (10^{-5})$	$1.8 \times (10^{-14})$
$\delta_{intercept}$	-7.7712	0.0409	$< 2 \times (10^{-16})$
δ_{zage}	-0.9999	0.0945	$< 2 \times (10^{-16})$
δ_{zage^2}	1.5207	0.0973	$< 2 \times (10^{-16})$
β_{sex}	0.4423	0.0225	$< 2 \times (10^{-16})$
β_{smoker}	0.4859	0.0270	$< 2 \times (10^{-16})$
$\beta_{zage\times smoker}$	0.3519	0.0296	$< 2 \times (10^{-16})$

Figures 8.10 and 8.11 show the smooth and crude death rates for the four combinations of gender and smoker status. It can be seen that until age 27 the mortality rates are slightly decreasing for non-smokers (see the upper graphs of Figures 8.10 and 8.11). We think that this is because of the accident hump which is observed for younger adults. This decrease is less obvious for smokers and lasts until age 18 (see the lower

graphs of Figures 8.10 and 8.11). However this effect might not be significant for smokers as there are very few data.

To see the effect of the age–smoker interaction in the model, we plot inception rates of non–smokers and the inception rates of smokers against age. This graph can be seen in Figure 8.12. The two lines cross at around age 25 and again, the lack of data below age 30 might have caused this effect.

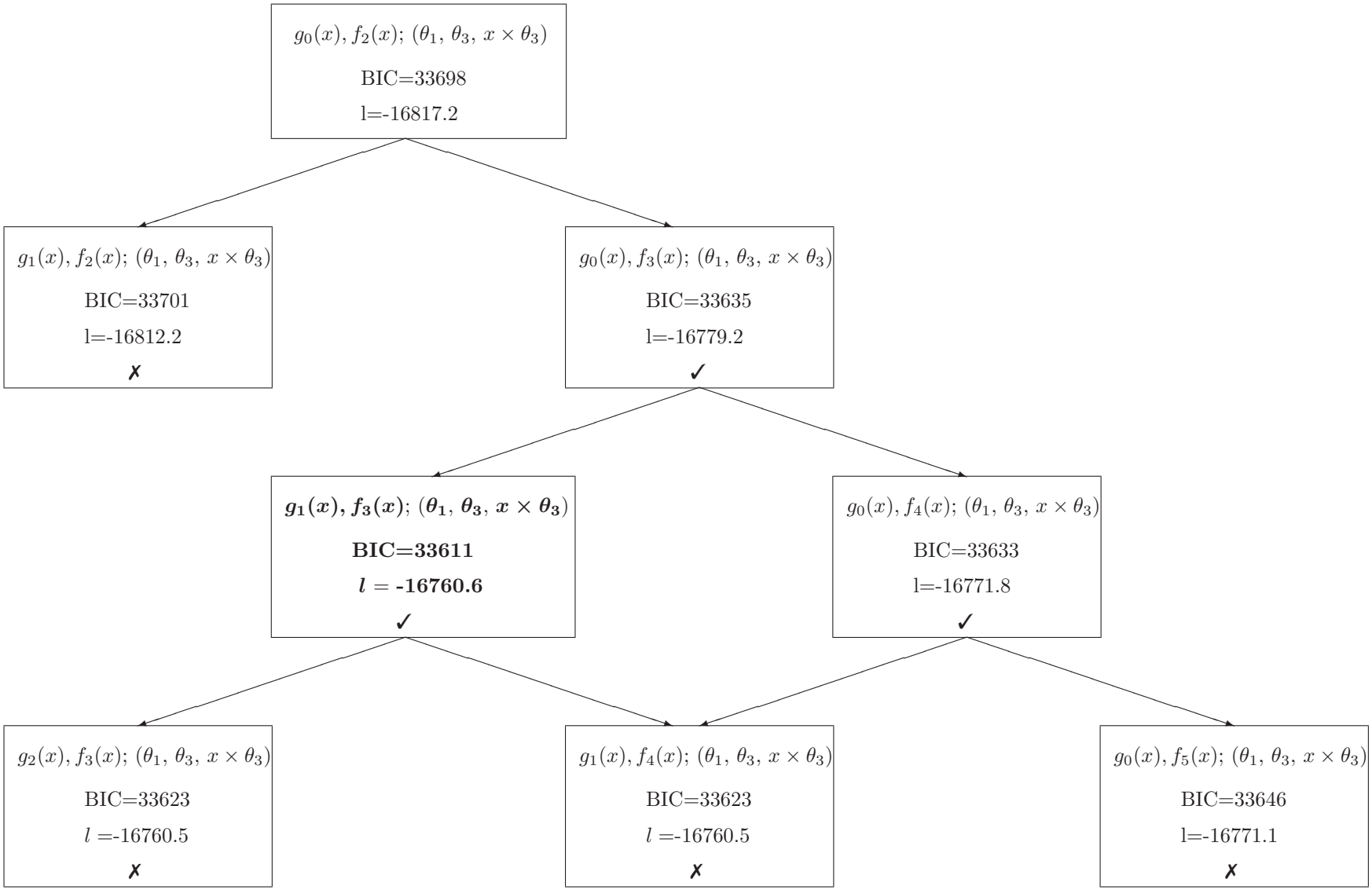


Figure 8.9: Model selection for death.

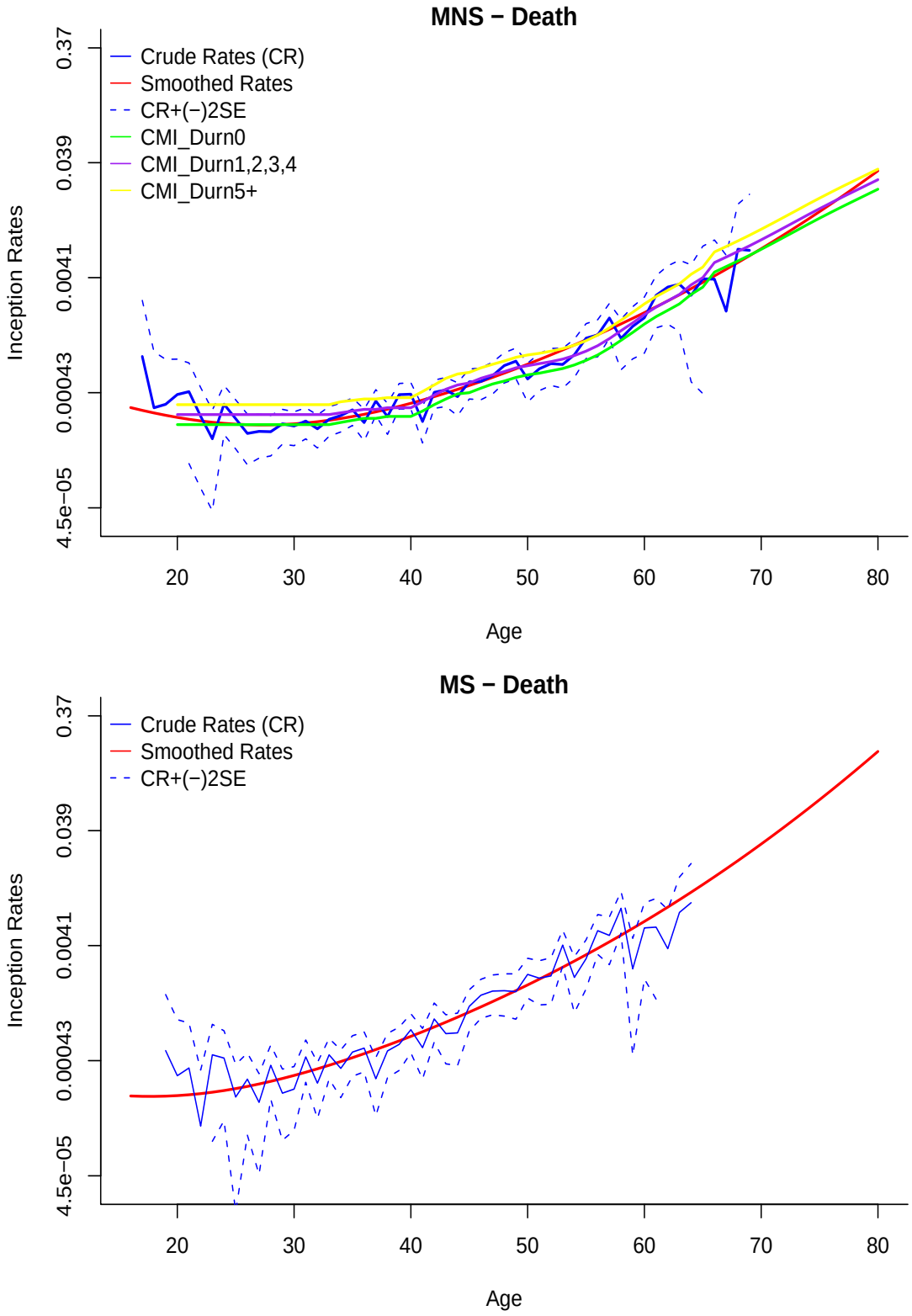


Figure 8.10: Graphs of diagnosis inception rates for death for males, non–smokers (MNS) and smokers (MS).

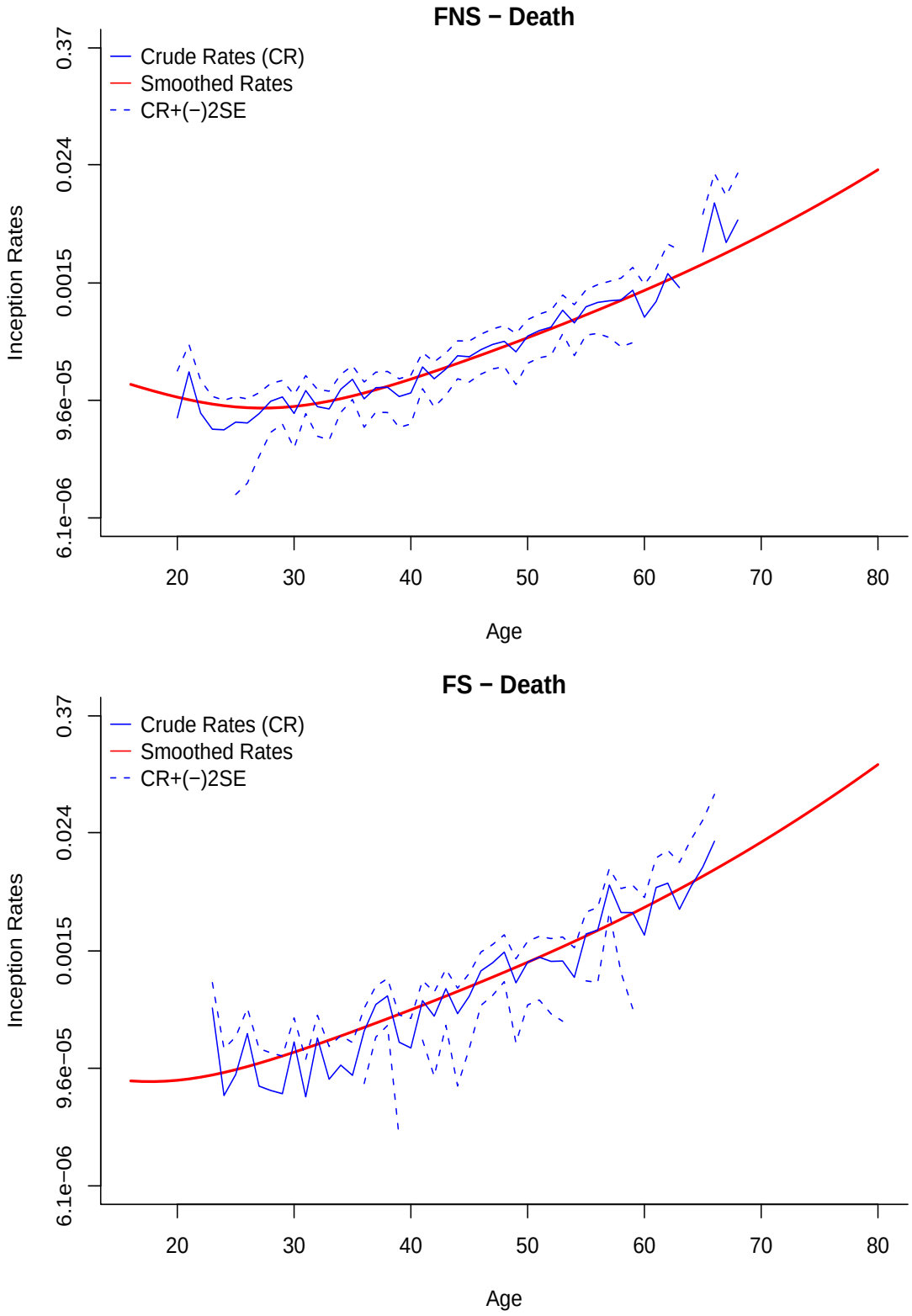


Figure 8.11: Graphs of diagnosis inception rates for death for females, non–smokers (FNS) and smokers (FS).

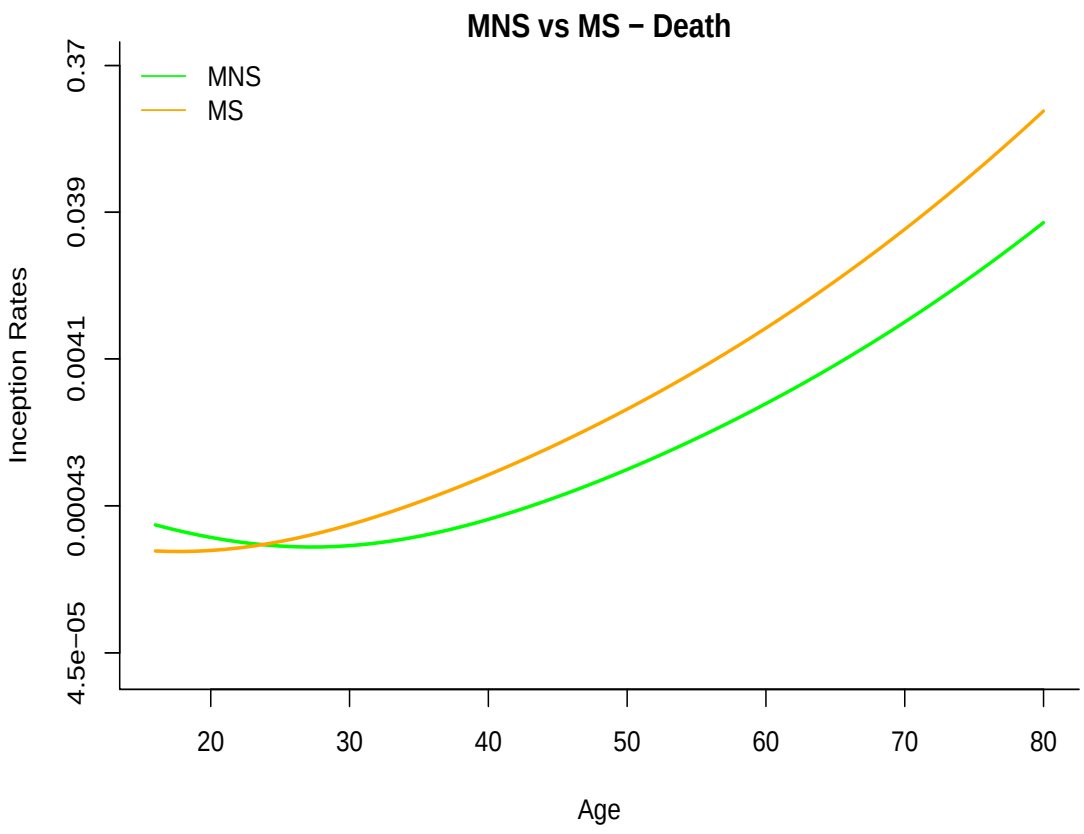


Figure 8.12: Graphs of diagnosis inception rates for death for males, non–smokers vs. smokers.

Heart attack

For heart attack inception rates, Figure 8.13 shows that the smallest BIC is obtained for the model with a quadratic age polynomial $(g_0(x), f_3(x))$. For each different age polynomial we have tried, sex and smoker status are found to be important. When the order of the polynomial is greater than 3, the age - smoker interaction stays in models as well. Our best model includes an age^2 - smoker interaction term.

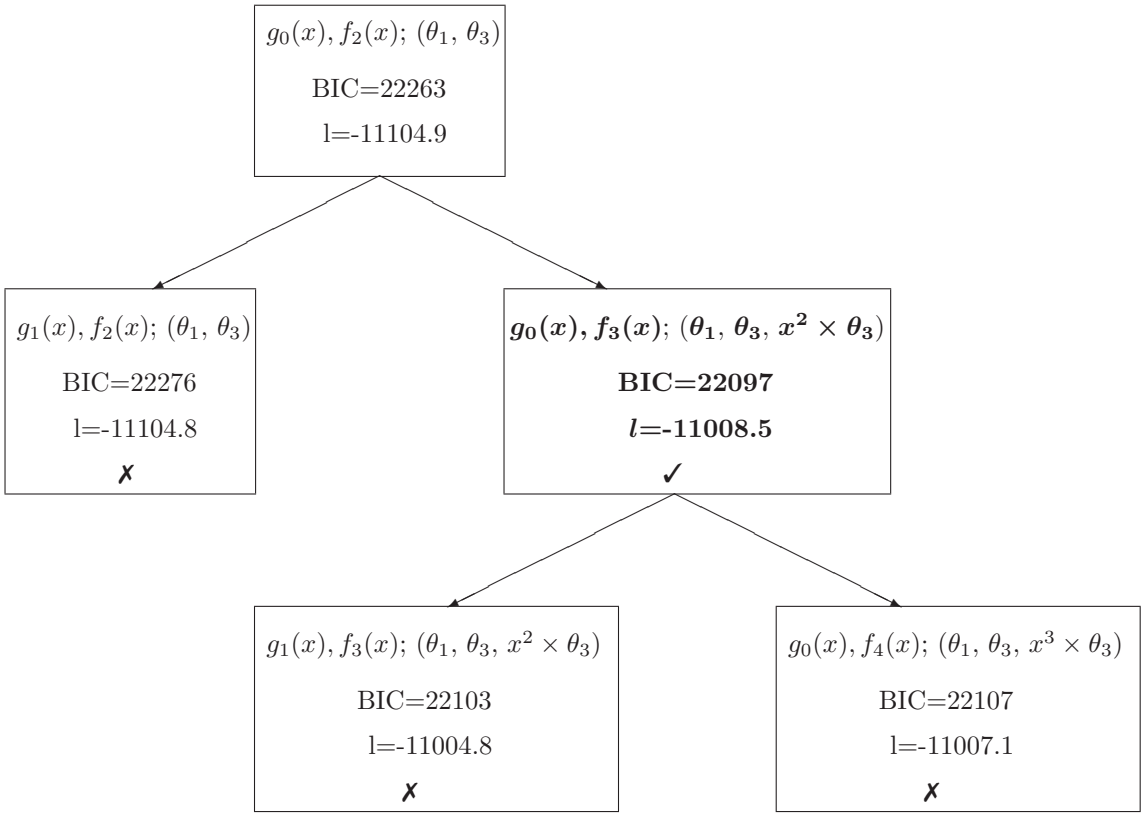


Figure 8.13: Model selection for heart attack.

The function we use to smooth the heart attack rates is given in (8.9), and parameter estimates are shown in Table 8.4. Males and smokers have higher heart attack rates. These two, i.e being a male and a smoker, are known to be the two major risk factors of cardiac diseases (see Chatterjee *et al.* (2009)).

$$\lambda^{HA} = \exp\left(\delta_{int} + \delta_{zage}x + \delta_{zage^2}x^2 + \beta_{sex}\theta_1 + \beta_{smoker}\theta_3 + \beta_{zage^2\times smoker}x^2 \times \theta_3\right) \tag{8.9}$$

Table 8.4: ML estimates of parameters under the best model for heart attack.

Parameter	Estimate	Std. Error	p-value
$\delta_{intercept}$	-11.1938	0.0895	$< 2 \times (10^{-16})$
δ_{zage}	5.4266	0.3232	$< 2 \times (10^{-16})$
δ_{zage^2}	-3.3297	0.2801	$< 2 \times (10^{-16})$
β_{sex}	1.9150	0.0750	$< 2 \times (10^{-16})$
β_{smoker}	1.5299	0.0586	$< 2 \times (10^{-16})$
$\beta_{zage^2\times smoker}$	-0.2798	0.0589	$2.1 \times (10^{-6})$

Figures 8.14 and 8.15 show the smooth and crude heart attack inception rates for the four combinations of gender and smoker status. The CMI produced heart attack inception rates for males, non-smokers for different policy durations. These rates are shown in the males and non-smokers graph (upper graph in Figure 8.14). There is a decrease above age around 65 because of the negative coefficients of age^2 and age^2 - smoker interaction. Generally for the inception rates of cardiac diseases a decrease with age is not expected. Since we have hardly any heart attack data after age 65, the inception rates after this age are based on extrapolation from the function and this might be the possible reason of this unexpected decrease after this age. For females we have less heart attack data compared to males (see Figure 8.15). Hence we could not provide the 'crude rates -2 standard errors' for most of the cases.

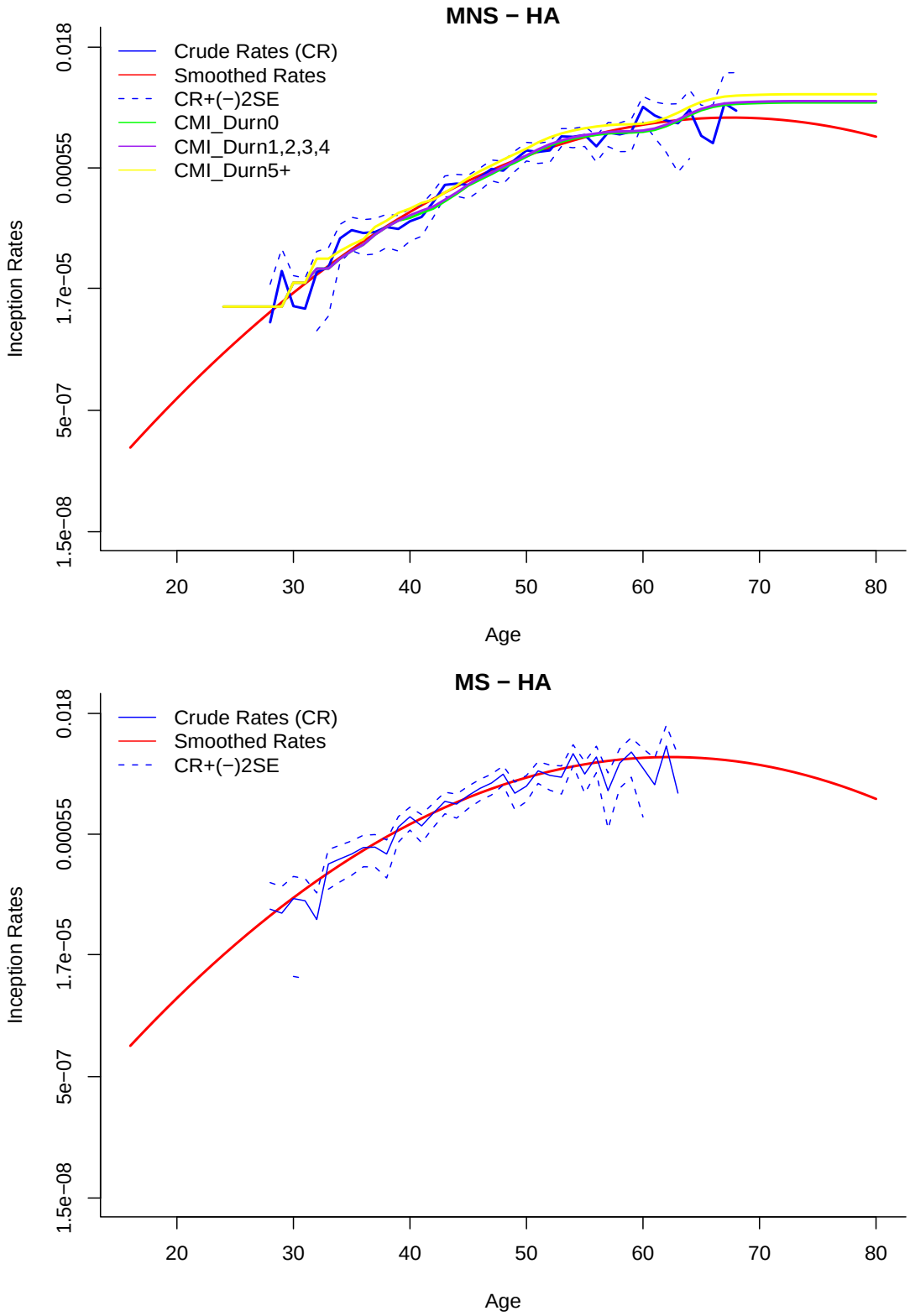


Figure 8.14: Graphs of diagnosis inception rates for heart attack for males, non–smokers (MNS) and smokers (MS).

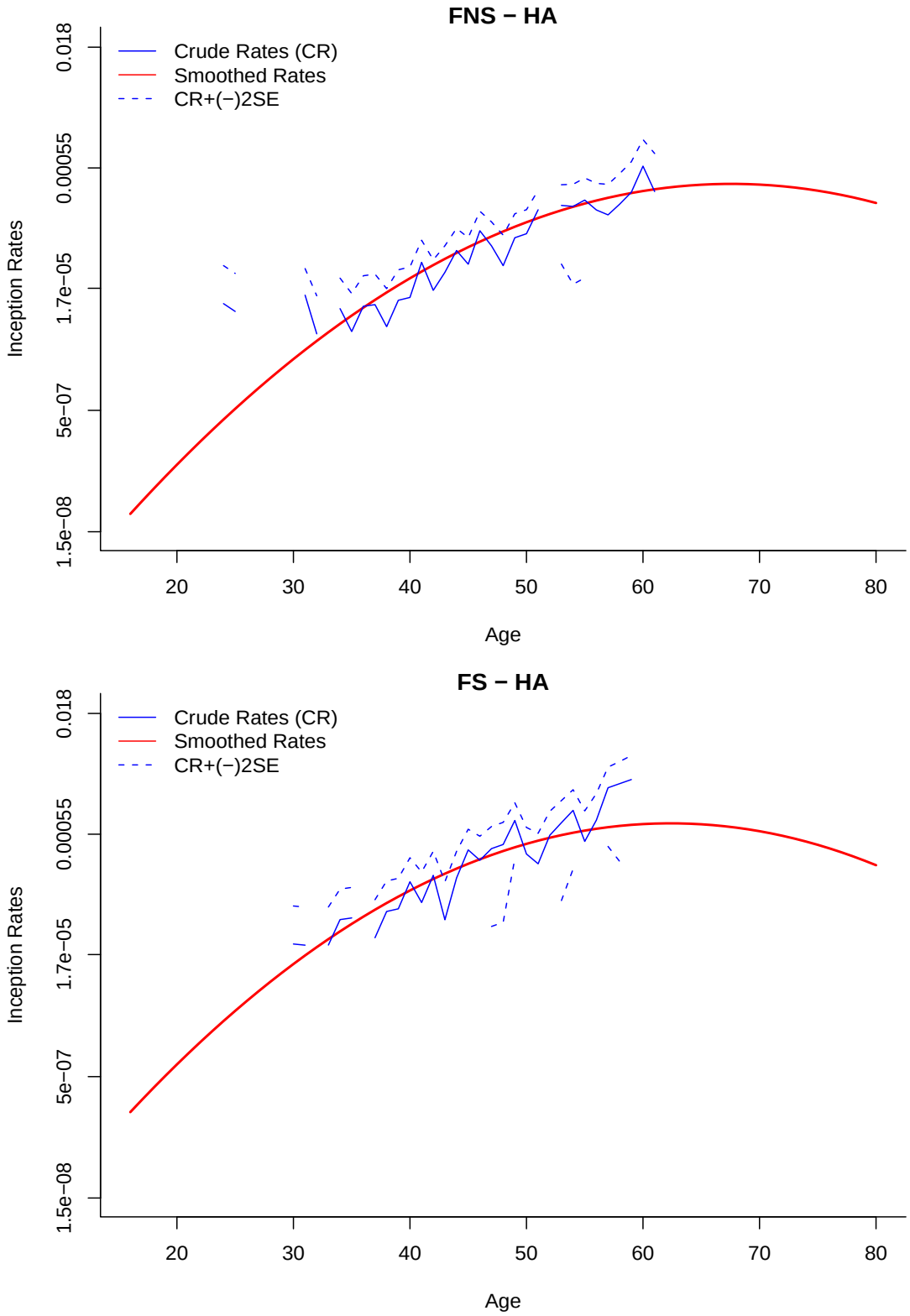


Figure 8.15: Graphs of diagnosis inception rates for heart attack for females, non–smokers (FNS) and smokers (FS).

Kidney failure

In the modelling of kidney failure, the exponential age term was not significant and was dropped from the model. Between the $g_0(x), f_1(x)$ and $g_1(x), f_1(x)$ models, the former is selected due to its lower BIC value. Sex is found to be the only significant covariate in smoothing the crude rates. Log-likelihood and BIC values of these models are given in Figure 8.16.

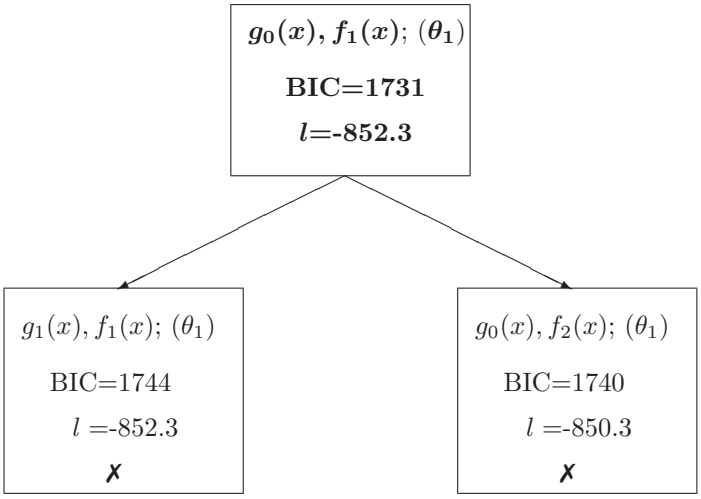


Figure 8.16: Model selection for kidney failure.

The smoothing function is shown in (8.10) and the estimates of the model parameters are given in Table 8.5. The positive coefficient for sex indicates a higher kidney failure rate for males.

$$\lambda^{KF} = \exp\left(\delta_{int} + \beta_{sex}\theta_1\right) \tag{8.10}$$

Table 8.5: ML estimates of parameters under the best model for kidney failure.

Parameter	Estimate	Std. Error	p-value
$\delta_{intercept}$	-12.7364	0.2582	$< 2 \times (10^{-16})$
β_{sex}	1.5805	0.2811	$1.9 \times (10^{-8})$

Figure 8.17 shows the crude and smoothed inception rates for kidney failure. Since we do not have many claims for this cause, especially for women, there is a large number of zero crude rates across ages. Thus the rates are given in the actual scale rather than the log scale.

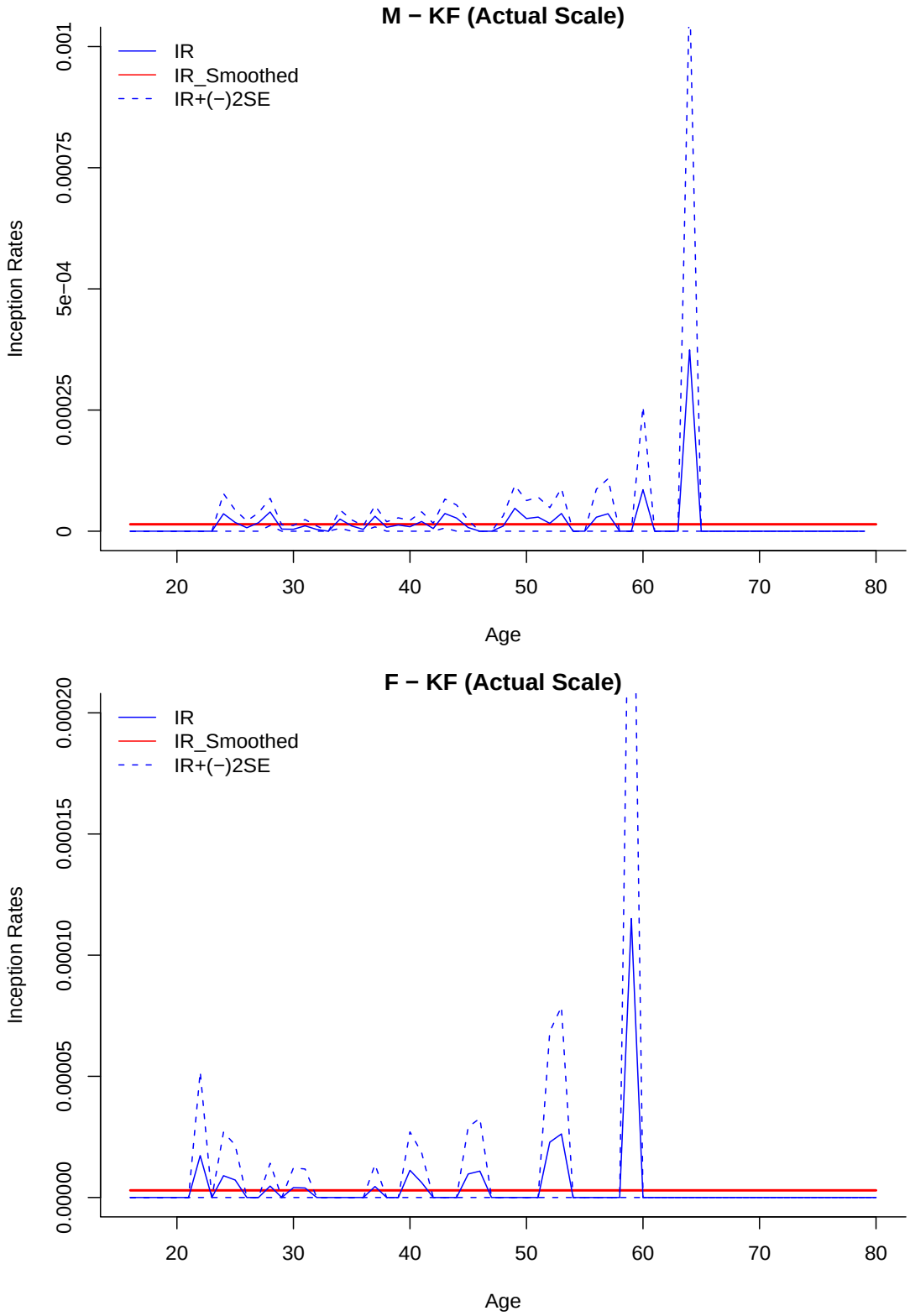


Figure 8.17: Graphs of diagnosis inception rates for kidney failure for males (M) and females (F).

Major organ transplantation

None of the considered covariates, including age, were significant in modelling the rates for major organ transplant. As can be seen from Figure 8.18, $g_0(x), f_1(x)$ has the lowest BIC value. Thus, for this cause we simply use an exponential intercept term in the modelling (see (8.11)) where the estimated coefficient is given in Table 8.6.

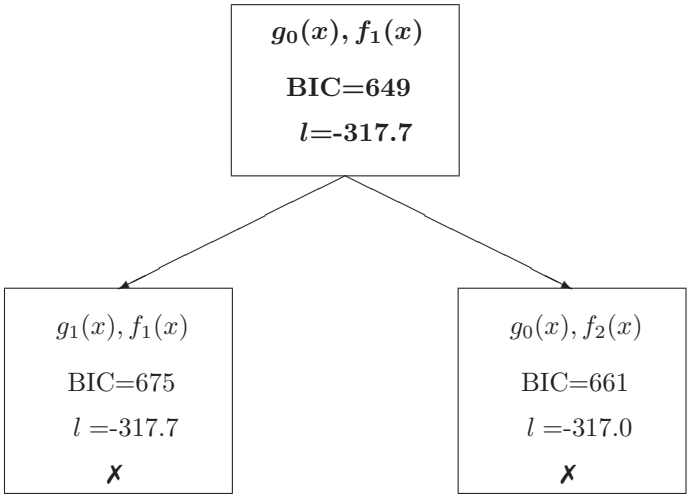


Figure 8.18: Model selection for major organ transplant.

$$\lambda^{MOT} = \exp(\delta_{int}) \tag{8.11}$$

Table 8.6: ML estimates of parameters under the best model for major organ transplantation.

Parameter	Estimate	Std. Error	p-value
$\delta_{intercept}$	-12.7501	0.1741	$< 2 \times (10^{-16})$

Figure 8.19 shows the crude and modelled MOT rates. The graph is given in the actual scale due to zero crude rates at various ages.

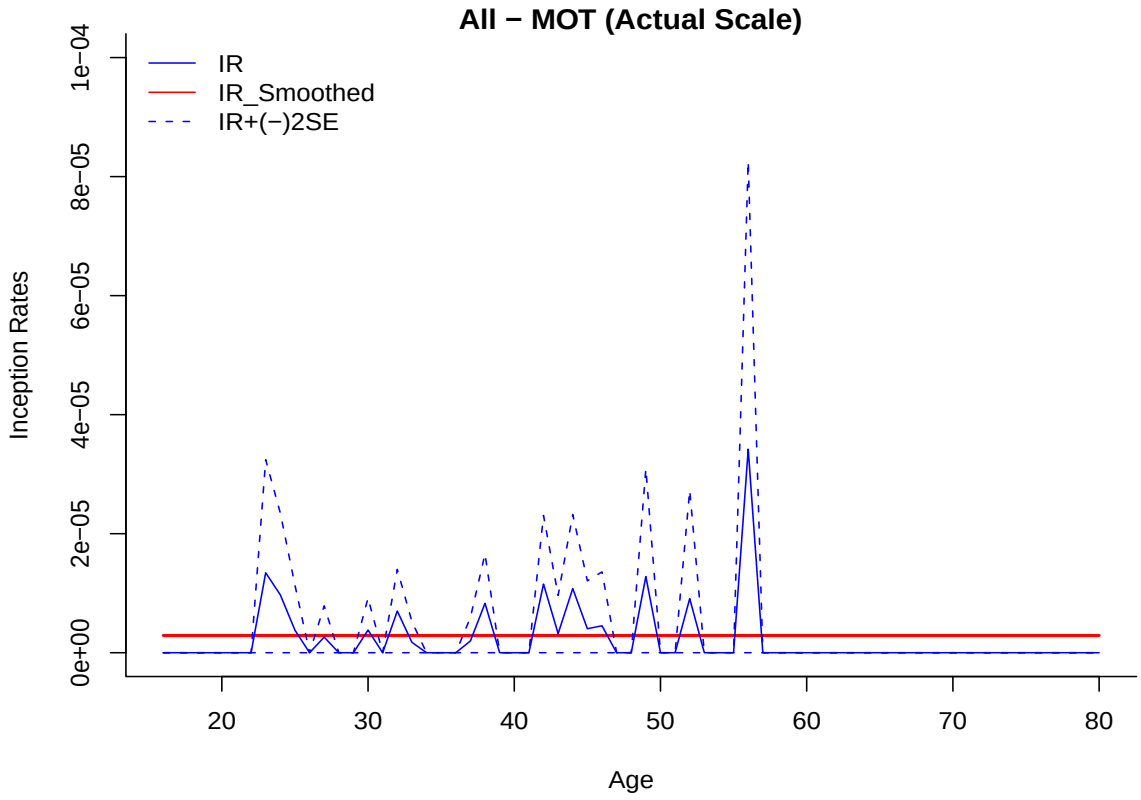


Figure 8.19: Graph of diagnosis inception rates for major organ transplant for all population.

Multiple sclerosis (MS)

We started searching for the best model for multiple sclerosis with a first order exponential age polynomial ($g_0(x), f_2(x)$). However the age term was found to be insignificant in the model and dropping it gave a model with lower BIC value ($g_0(x), f_1(x)$). Including a linear intercept in the model did not improve the model fit and we ended up using a $g_0(x), f_1(x)$ function. The log-likelihood and BIC values for these models are given in Figure 8.20. The effects of sex, smoking status and policy duration on multiple sclerosis were found to be important.

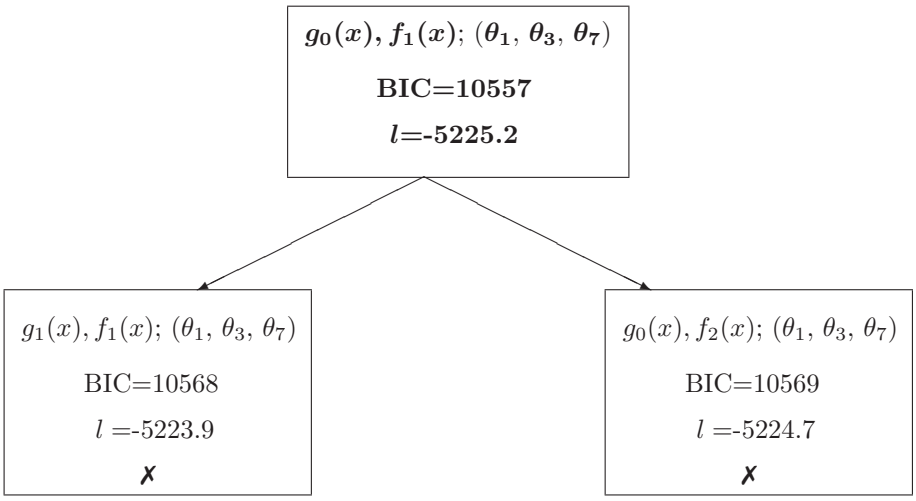


Figure 8.20: Model selection for multiple sclerosis.

The function we use to smooth the multiple sclerosis rates is given in (8.12) and parameter estimates are shown in Table 8.7. As can be seen from the table, males have lower MS rates than females and smokers have higher rates than non-smokers. In Chapter 3 we mentioned that there is a 6-months waiting period for this disease before a claim is admitted. We can see the effect of this waiting period on the coefficient of the first year policy duration: it is considerably lower than for the other durations.

$$\lambda^{MS} = \exp\left(\delta_{int} + \beta_{sex}\theta_1 + \beta_{smoker}\theta_3 + \beta_{poldur}\theta_7\right) \tag{8.12}$$

Table 8.7: ML estimates of parameters under the best model for multiple sclerosis.

Parameter	Estimate	Std. Error	p-value
$\delta_{intercept}$	-9.2773	0.0508	$< 2 \times (10^{-16})$
β_{sex}	-0.7460	0.0760	$< 2 \times (10^{-16})$
β_{smoker}	0.3935	0.0847	$3.4 \times (10^{-6})$
$\beta_{poldur0}$	-0.8727	0.0944	$< 2 \times (10^{-16})$
$\beta_{poldur1}$	-0.0607	0.0732	0.4069
$\beta_{poldur2}$	0.1876	0.0746	0.0119
$\beta_{poldur3}$	0.2174	0.0873	0.0127
$\beta_{poldur4}$	0.1512	0.1092	0.1661
$\beta_{poldur5+}$	0.3772	0.0750	$4.9 \times (10^{-7})$

Figures 8.21 and 8.22 show the crude and modelled multiple sclerosis rates for the four combinations of gender and smoking status. These rates (both crude and modelled) are the weighted averages of policy durations. Inception rates can be obtained for an average policy duration by weighting the duration-specific inception rates with duration-specific exposures in a similar way as in (7.8) and (8.7). In the figures we also show inception rates for specific durations, Duration 0 and Duration 5+. The reason for the weighted rates being close to rates with policy duration 5+ is that most of the MS claims have long policy durations. Most of the lower bounds of the confidence intervals for the crude rates could not be shown for smokers as there are very few data (the lower graphs in Figures 8.21 and 8.22).

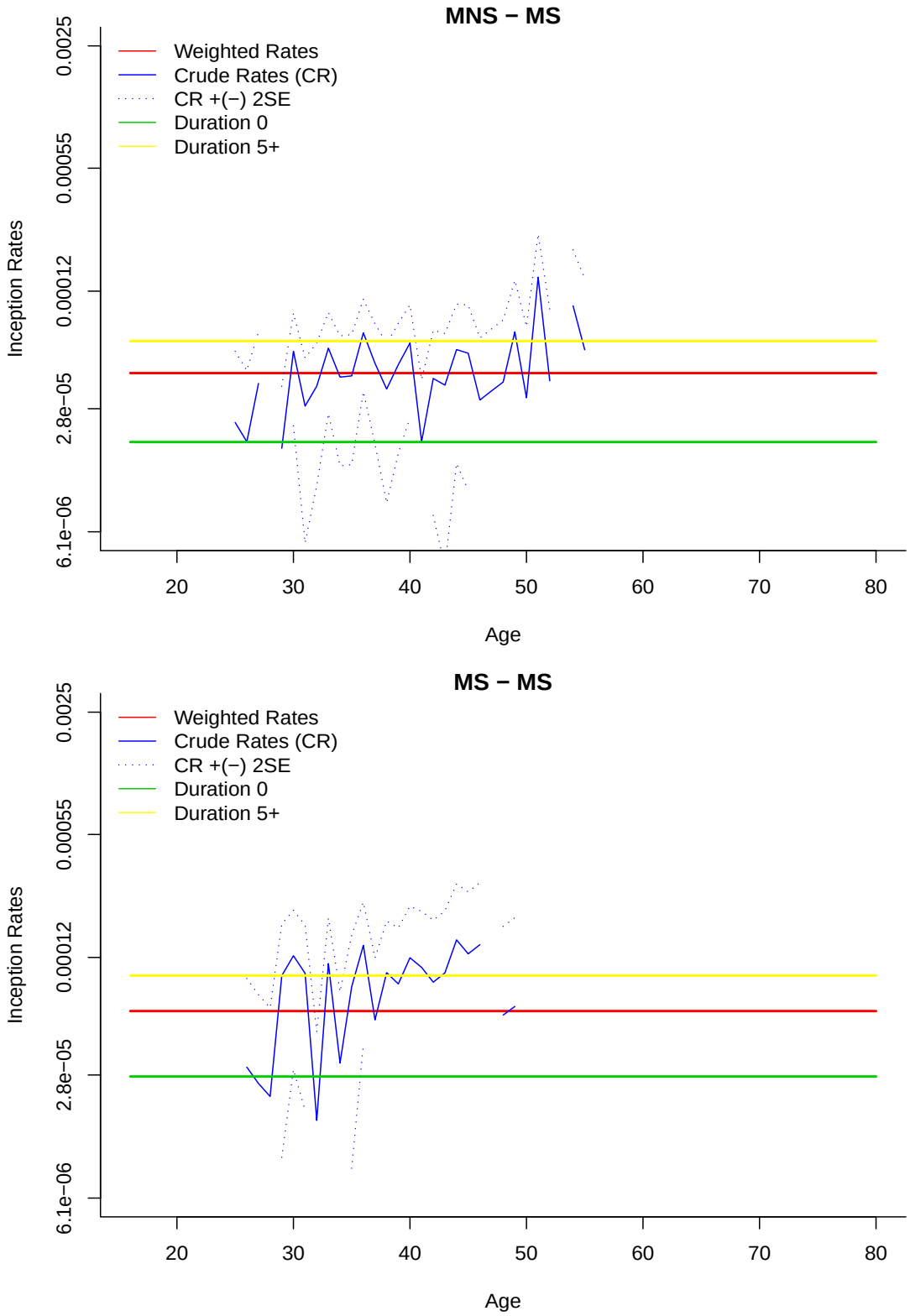


Figure 8.21: Graphs of diagnosis inception rates for multiple sclerosis for males, non-smokers (MNS) and smokers (MS).

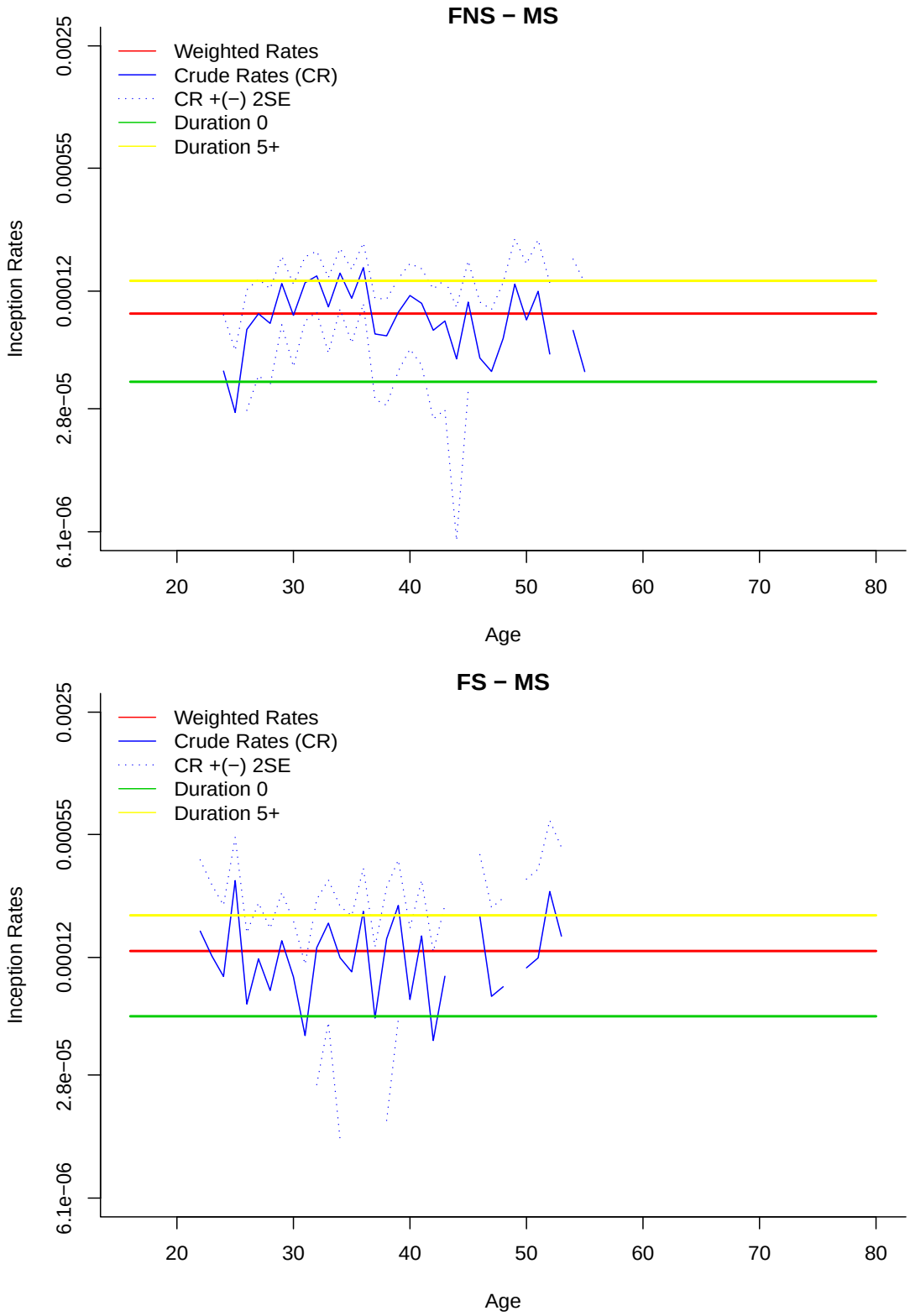


Figure 8.22: Graphs of diagnosis inception rates for multiple sclerosis for females, nonsmokers (FNS) and smokers (FS).

Other Causes

In modelling other causes, the model with an exponential first degree age polynomial $(g_0(x), f_2(x))$ has a better fit compared to the model with a quadratic age polynomial $(g_0(x), f_3(x))$ or compared to the model which includes a linear intercept term $(g_1(x), f_2(x))$. BIC values can be seen in Figure 8.23.

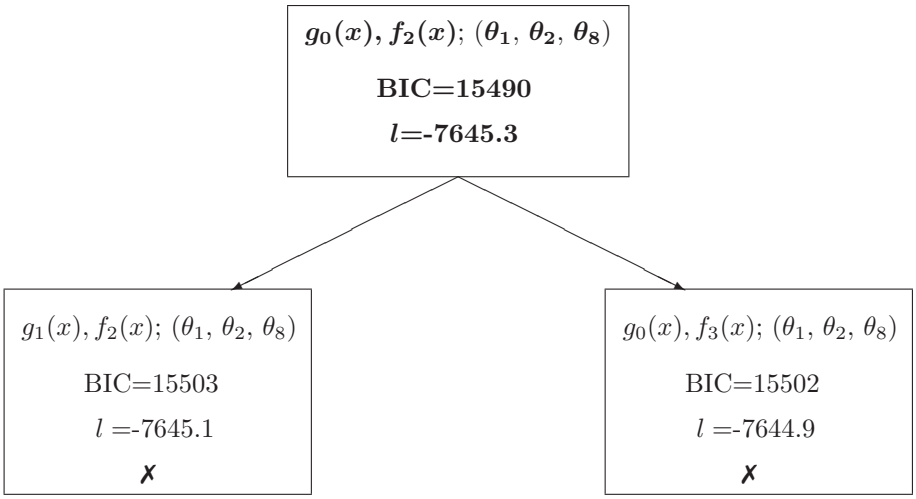


Figure 8.23: Model selection for other causes.

The best model includes sex, benefit type and office covariates and is given by equation (8.13). For Office 7 we do not have any other-cause claims. Therefore, the analysis is performed excluding this office from the analysis. The estimated parameters of the model are presented in Table 8.8. Other-cause rates are higher for males and stand alone policies. The coefficient of Office 1 is the median of office coefficients. Office 6 has the highest coefficient and thus claims from this office are associated with the highest rates and those from Office 12 with the lowest rates.

$$\lambda^{Other} = \exp\left(\delta_{int} + \delta_{zage}x + \beta_{sex}\theta_1 + \beta_{bentype}\theta_2 + \beta_{office}\theta_8\right) \tag{8.13}$$

Table 8.8: ML estimates of parameters under the best model for other diseases.

Parameter	Estimate	Std. Error	p-value
$\delta_{intercept}$	-9.7664	0.0727	$< 2 \times (10^{-16})$
δ_{zage}	0.8427	0.0371	$< 2 \times (10^{-16})$
β_{sex}	0.6769	0.0641	$< 2 \times (10^{-16})$
$\beta_{bentype}$	0.3822	0.0752	$3.7 \times (10^{-7})$
$\beta_{office1}$	0.0054	0.1028	0.9579
$\beta_{office2}$	0.4897	0.0773	$< 2.4 \times (10^{-10})$
$\beta_{office3}$	-0.4387	0.3274	0.1802
$\beta_{office4}$	0.4308	0.1778	0.0154
$\beta_{office5}$	-0.4789	0.2345	0.0411
$\beta_{office6}$	0.8111	0.2077	$< 9.4 \times (10^{-5})$
$\beta_{office7}$	-	-	-
$\beta_{office8}$	0.6979	0.0747	$< 2 \times (10^{-16})$
$\beta_{office9}$	0.1077	0.1074	0.3160
$\beta_{office10}$	-0.6450	0.2113	0.0023
$\beta_{office11}$	0.0827	0.0821	0.3138
$\beta_{office12}$	-1.0341	0.1704	$< 1.3 \times (10^{-9})$
$\beta_{office13}$	-0.0285	0.1635	0.8614

Figures 8.24 and 8.25 show crude and smooth rates for other causes for the combinations of gender and benefit type. These rates are the weighted average over the offices (calculated in a similar way as in (7.8)). We also show the rates for the individual offices (Office 12, Office 1 and Office 6) in the figures. We are using office specific exposures to weight the rates. It is mentioned previously that we have very few and irregular exposures below age 20 and above age 65 for individual offices and therefore we observe lack of smoothness below age 20 and above age 65 when we weight the rates. Here, for older ages we fix the exposure at age 65 as explained in Chapter 7. Although the same can be applied to younger ages, we have not fixed the exposure below age 20. Therefore, in the graphs it is seen that there is lack of smoothness in the weighted smoothed rates below age 20.

Especially for full accelerated policies, the weight of Office 1 for older ages is high (see upper graphs in Figures 8.24 and 8.25). Therefore the weighted rates become almost the same as for Office 1. For male, stand alone policies, on the other hand, other offices are also contributing and thus the weighted rates do not approach the rates for Office 1 (see the lower graph in Figure 8.24). Note that we could not provide any of the lower bounds of the confidence interval for female stand alone policies as we have very limited observed numbers of claims (see the lower graph in Figure 8.25).

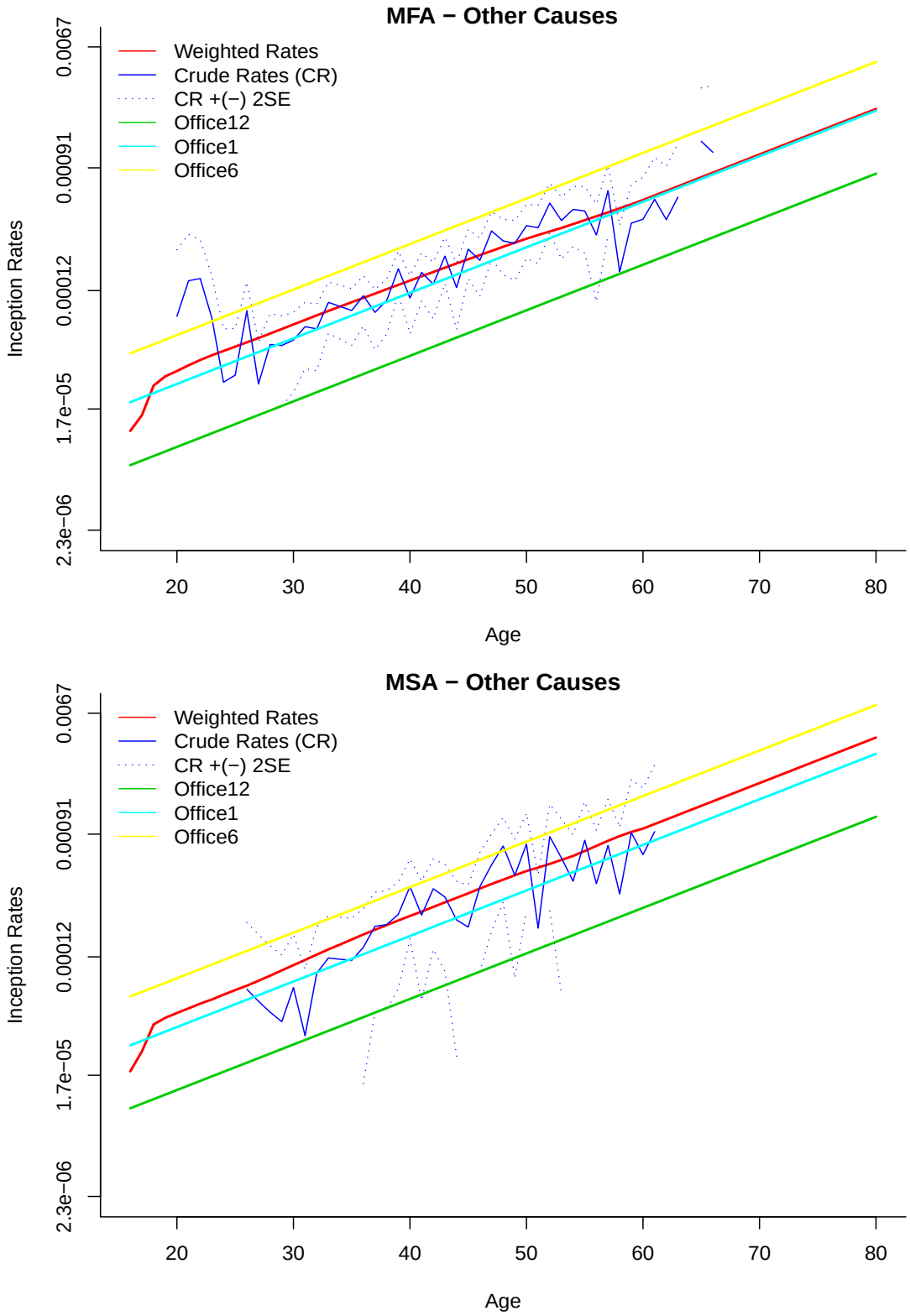


Figure 8.24: Graphs of diagnosis inception rates for other diseases for males, full accelerated (MFA) and stand alone policies (MSA).

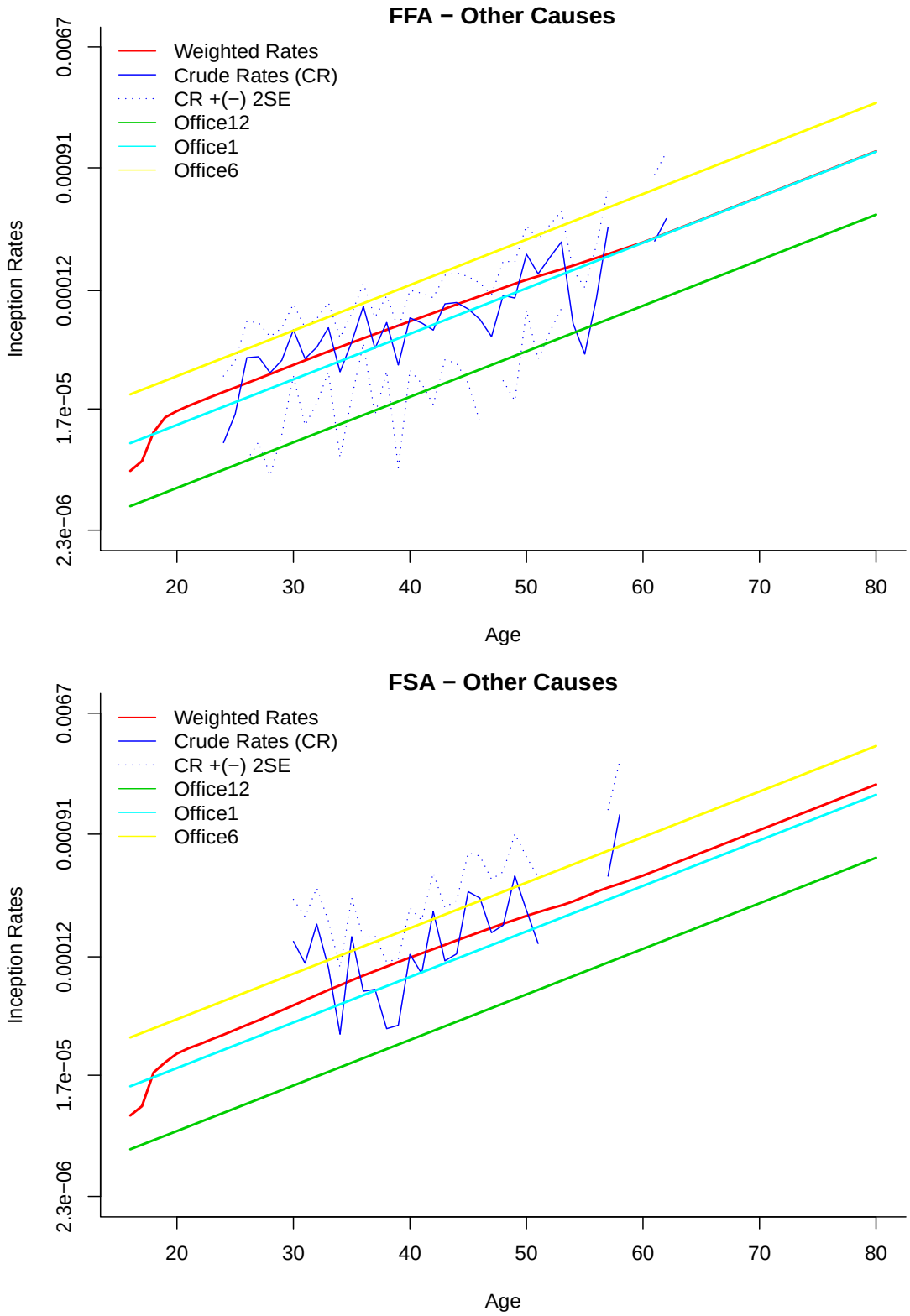


Figure 8.25: Graphs of diagnosis inception rates for other diseases for females, full accelerated (FFA) and stand alone policies (FSA).

Stroke

Stroke is modelled by using a first order exponential age polynomial ($g_0(x)$, $f_2(x)$) due to its lower BIC value compared to the other models considered. These models are shown in Figure 8.26 together with the best covariates under different age functions and their log-likelihood and BIC values.

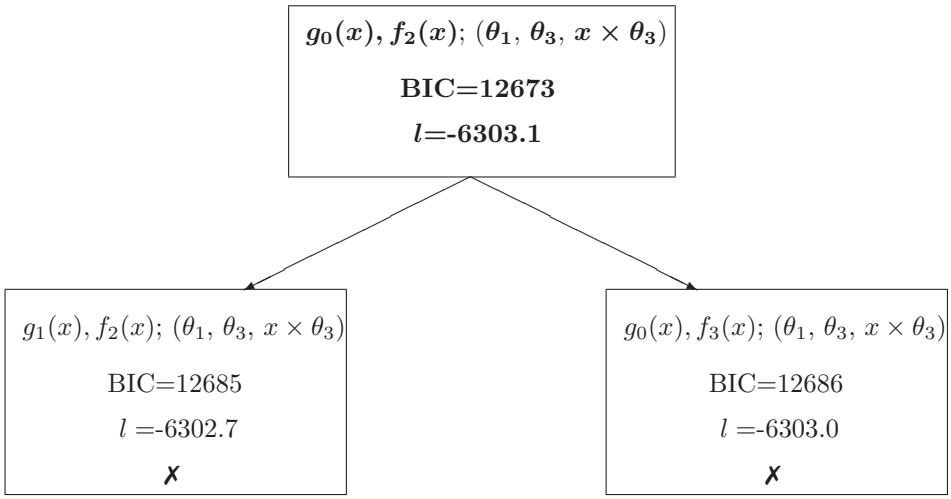


Figure 8.26: Model selection for stroke.

The selected model includes sex, smoker status and age - smoker interaction. The smoothing function is given in (8.14) and the estimated parameters are presented in Table 8.9. The positive coefficient for sex indicates that stroke rates are higher for males. Being a smoker also increases the inception rates. This effect increases with increasing age.

$$\lambda^{Stroke} = \exp\left(\delta_{int} + \delta_{zage}x + \beta_{sex}\theta_1 + \beta_{smoker}\theta_3 + \beta_{zage\times smoker}x \times \theta_3\right) \tag{8.14}$$

Table 8.9: ML estimates of parameters under the best model for stroke.

Parameter	Estimate	Std. Error	p-value
$\delta_{intercept}$	-9.7228	0.0605	$< 2 \times (10^{-16})$
δ_{zage}	1.0440	0.0467	$< 2 \times (10^{-16})$
β_{sex}	0.3764	0.0684	$3.1 \times (10^{-10})$
β_{smoker}	0.5236	0.0832	$3.8 \times (10^{-8})$
$\beta_{zage \times smoker}$	0.4137	0.0845	$9.8 \times (10^{-7})$

Figures 8.27 and 8.28 show the crude and smooth stroke rates for the subsets of gender and smoker status. In the males, non-smokers plot, rates produced by the CMI are also included. The CMI rates are fixed at age 60, due to lack of data after this age.

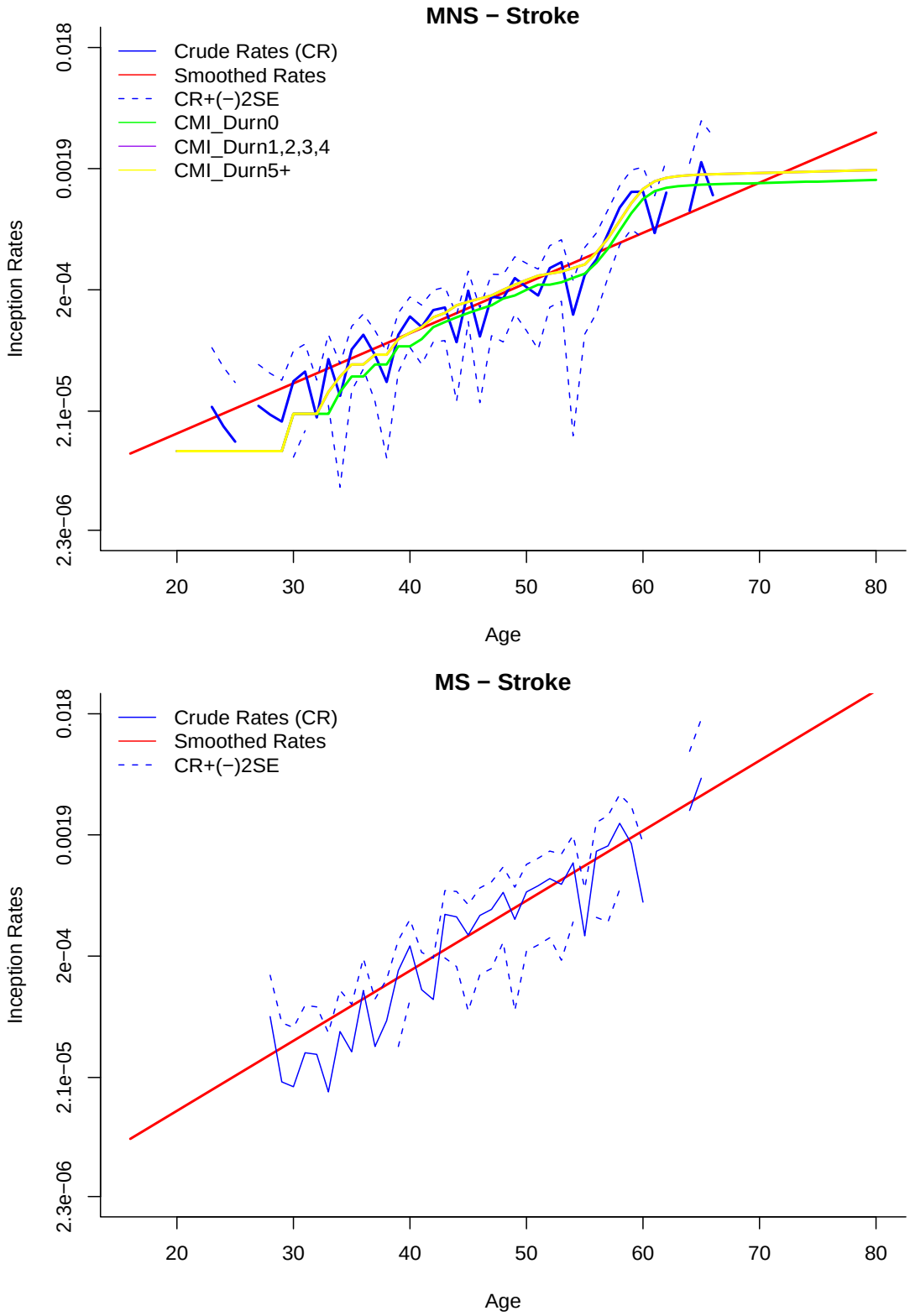


Figure 8.27: Graphs of diagnosis inception rates for stroke for males, nonsmokers (MNS) and smokers (MS).

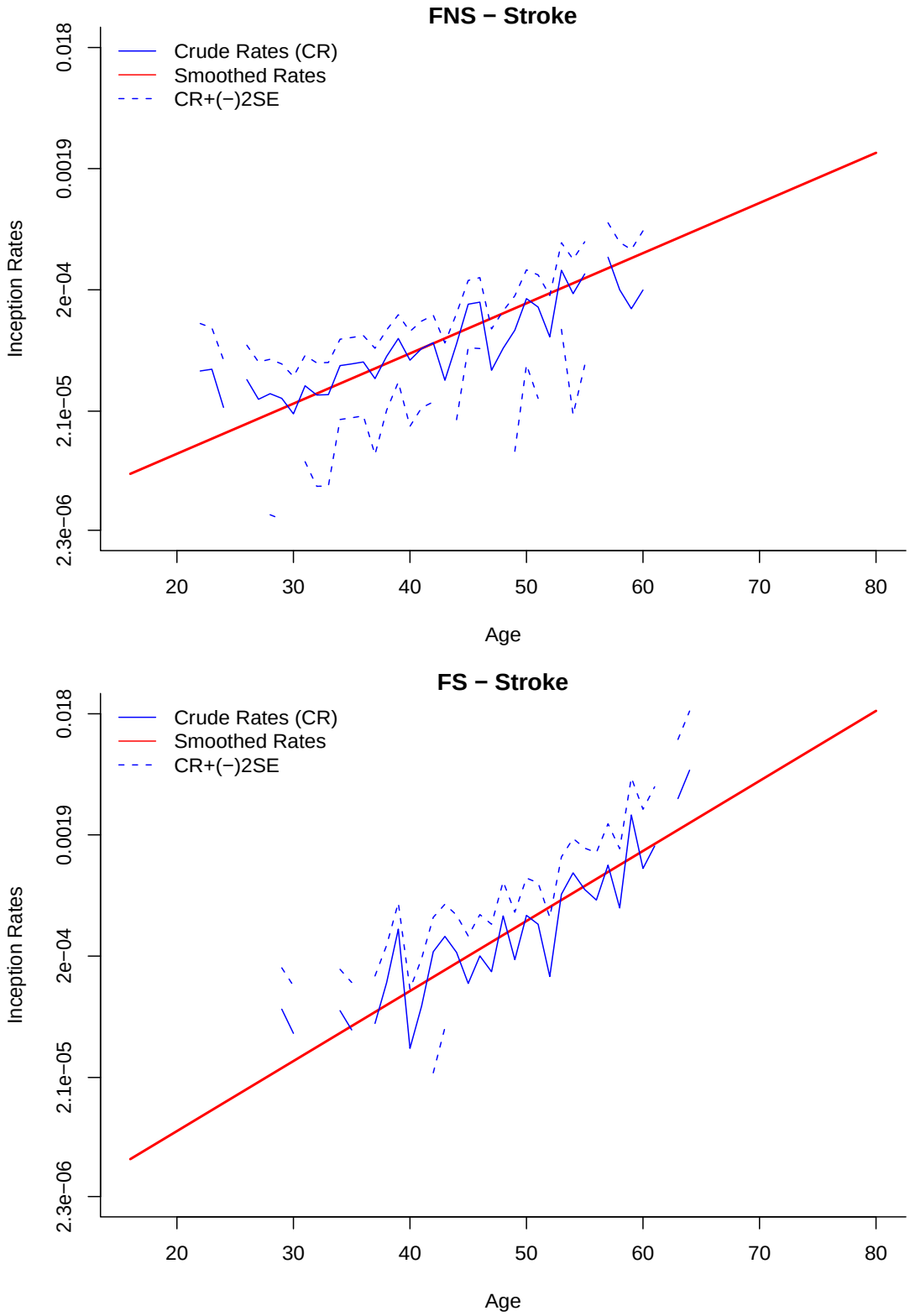


Figure 8.28: Graphs of diagnosis inception rates for stroke for females, nonsmokers (FNS) and smokers (FS).

The effect of the interaction term can be seen in Figure 8.29 which shows the non–smoker rates against the smoker rates. These two lines intersect at about age 25 meaning that the fit of the model is poor below age 25. As stated before, the possible reason for this is extrapolating from the function when the data is sparse.

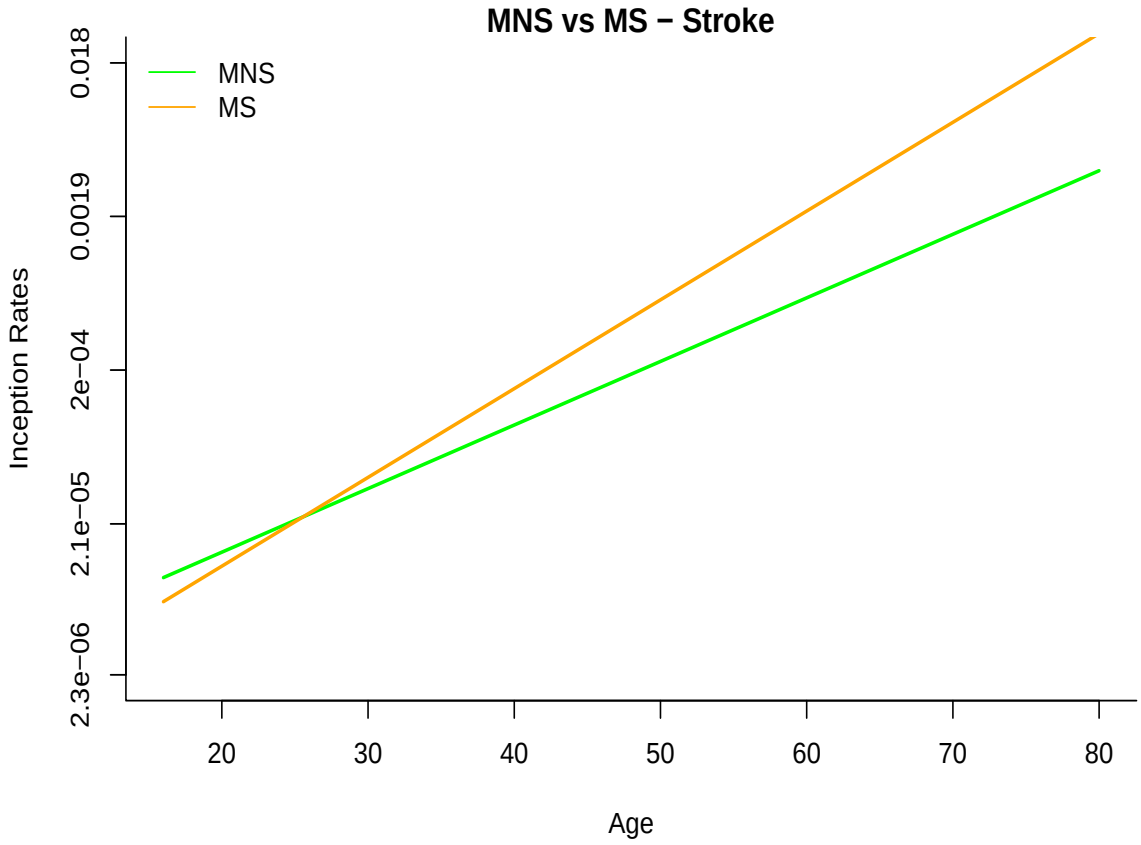


Figure 8.29: Graphs of diagnosis inception rates for stroke for males, nonsmokers vs smokers.

Total and permanent disability (TPD)

In terms of BIC, the model with age function $g_0(x), f_2(x)$ is found to give a better fit than the models with $g_0(x), f_3(x)$ and $g_1(x), f_2(x)$ age functions (see Figure 8.30). The selected models under different age functions include year and policy duration.

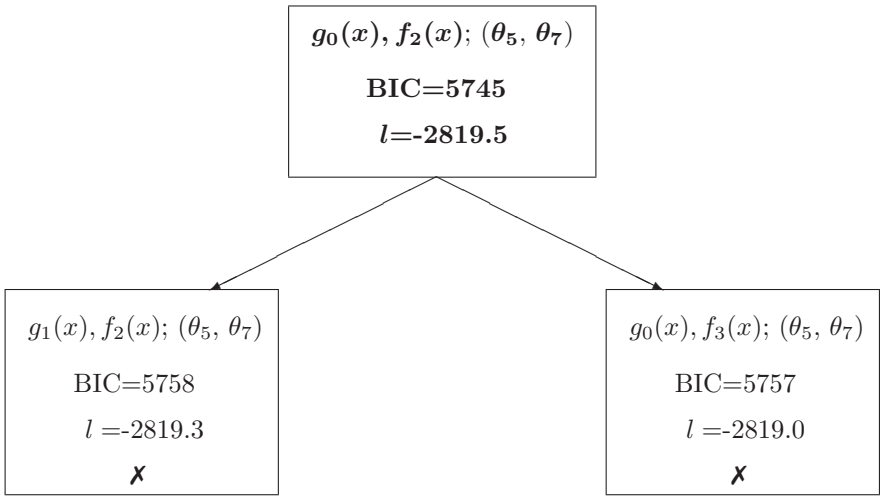


Figure 8.30: Model selection for TPD.

The smoothing function is given in (8.15) and the estimated parameters are shown in Table 8.10. For TPD claims, insurance companies generally wait at least 6-months before the settlement to make sure the effect of the cause is permanent. This period may change for different causes of TPD (e.g. mental illness, accidents, etc.) and each office applies its own rules to accept and settle policies as TPD claims. Moreover, due to imprecise definition of TPD, the recorded date of diagnosis is not consistently determined across offices and across TPD claims.

$$\lambda^{TPD} = \exp\left(\delta_{int} + \delta_{zage}x + \beta_{zyear}\theta_5 + \beta_{poldur}\theta_7\right) \tag{8.15}$$

Table 8.10: ML estimates of parameters under the best model for total and permanent disability.

Parameter	Estimate	Std. Error	p-value
$\delta_{intercept}$	-10.3321	0.0637	$< 2 \times (10^{-16})$
δ_{zage}	0.8815	0.0630	$< 2 \times (10^{-16})$
β_{zyear}	-0.2564	0.0553	$3.6 \times (10^{-6})$
$\beta_{poldur0}$	-1.1459	0.1753	$6.3 \times (10^{-11})$
$\beta_{poldur1}$	-0.6855	0.1513	$5.9 \times (10^{-6})$
$\beta_{poldur2}$	0.1294	0.1206	0.2830
$\beta_{poldur3}$	0.0650	0.1415	0.6460
$\beta_{poldur4}$	0.6046	0.1319	$4.6 \times (10^{-6})$
$\beta_{poldur5+}$	1.0324	0.0886	$< 2 \times (10^{-16})$

Figures 8.31 to 8.33 show the crude and modelled TPD rates for different policy durations. These rates are weighted averages over years (see (8.7)). The effect of lack of exposure on the weighted smoothed inception rates can be seen in the values before age 20. Individual years (1999 and 2005) are also shown in these graphs. Finally, the modelled rates by CMI are included in the figures. We should mention that the CMI rates are for males and non-smokers and they are the same for policy durations 1 to 4. Therefore the CMI rates shown in the lower graph in Figure 8.31 are considerably higher than the crude rates. Note that the CMI rates are provided until age 65 for this cause. This is because most of the insurance companies restrict the age to 65 for TPD claims. However this is not a uniform practice and different companies use different upper limits. In our model, we extend the rates until age 80 as for the other diseases.

Because of the waiting period for this cause as explained above, there are almost no cases in the first year of the policy (see the upper graph in Figure 8.31). The number of claims are considerably higher for the long policy durations, and therefore only in the lower graph in Figure 8.33 are we able to give the lower bound of the confidence interval for the crude rates. This is because the acceptance of a TPD claim as a valid claim is subjective and changes from office to office and claim to claim. As explained earlier, the ABI set some standards for the definition of this disease in 2010, however the experience used in this study does not cover these adjustments.

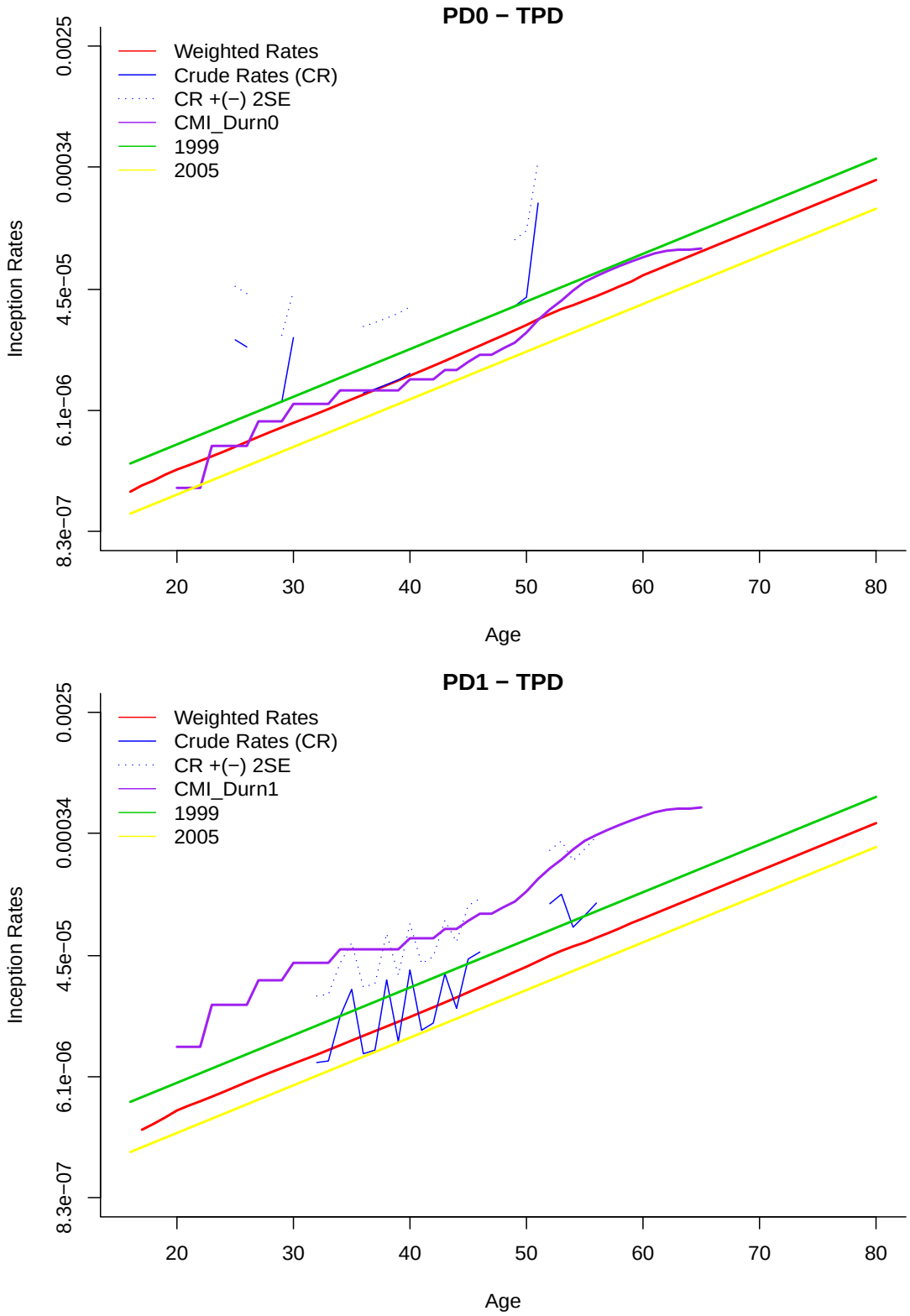


Figure 8.31: Graphs of diagnosis inception rates for TPD for policy durations 0 (PD0) and 1 (PD1).

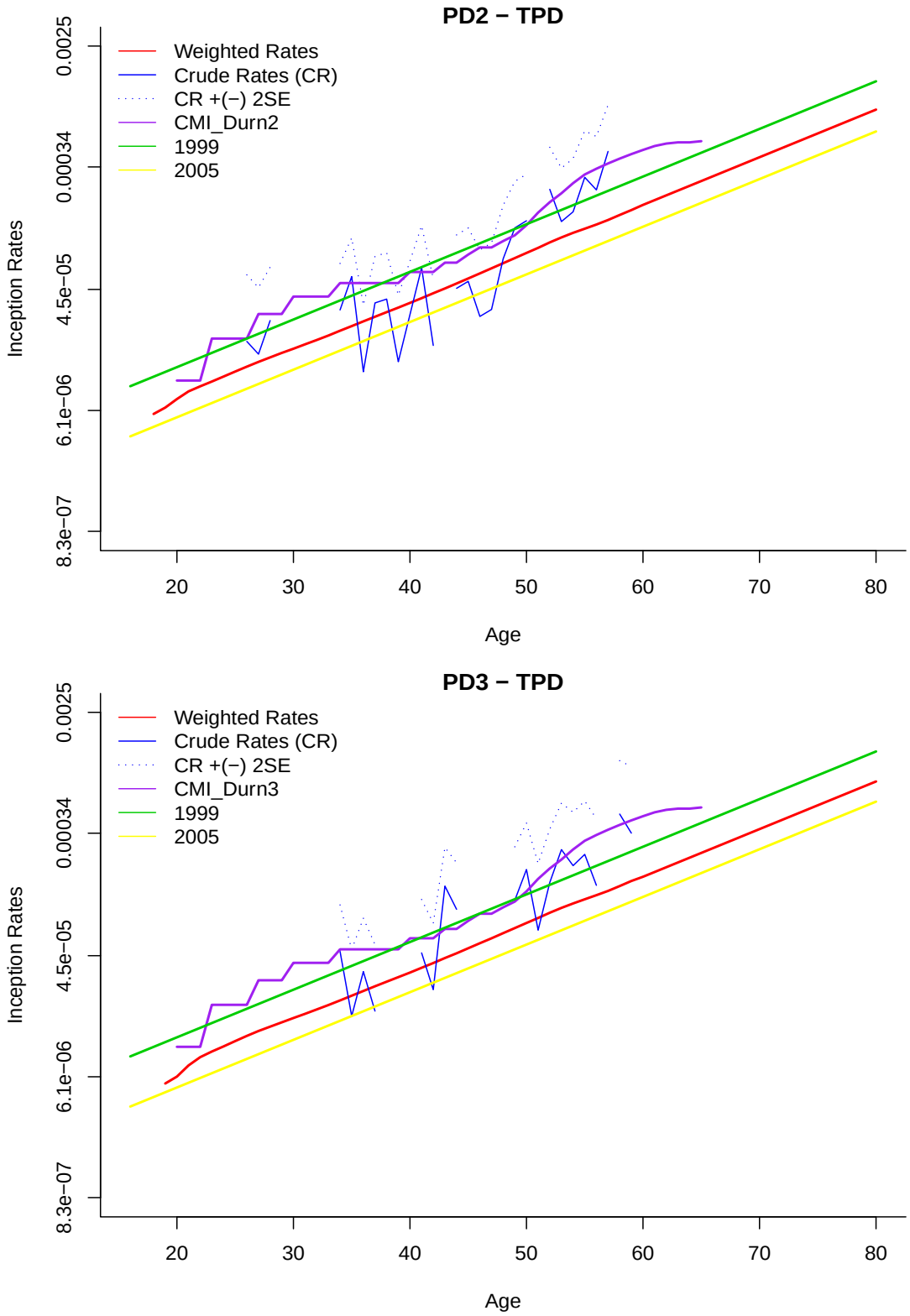


Figure 8.32: Graphs of diagnosis inception rates for TPD for policy durations 2 (PD2) and 3 (PD3).

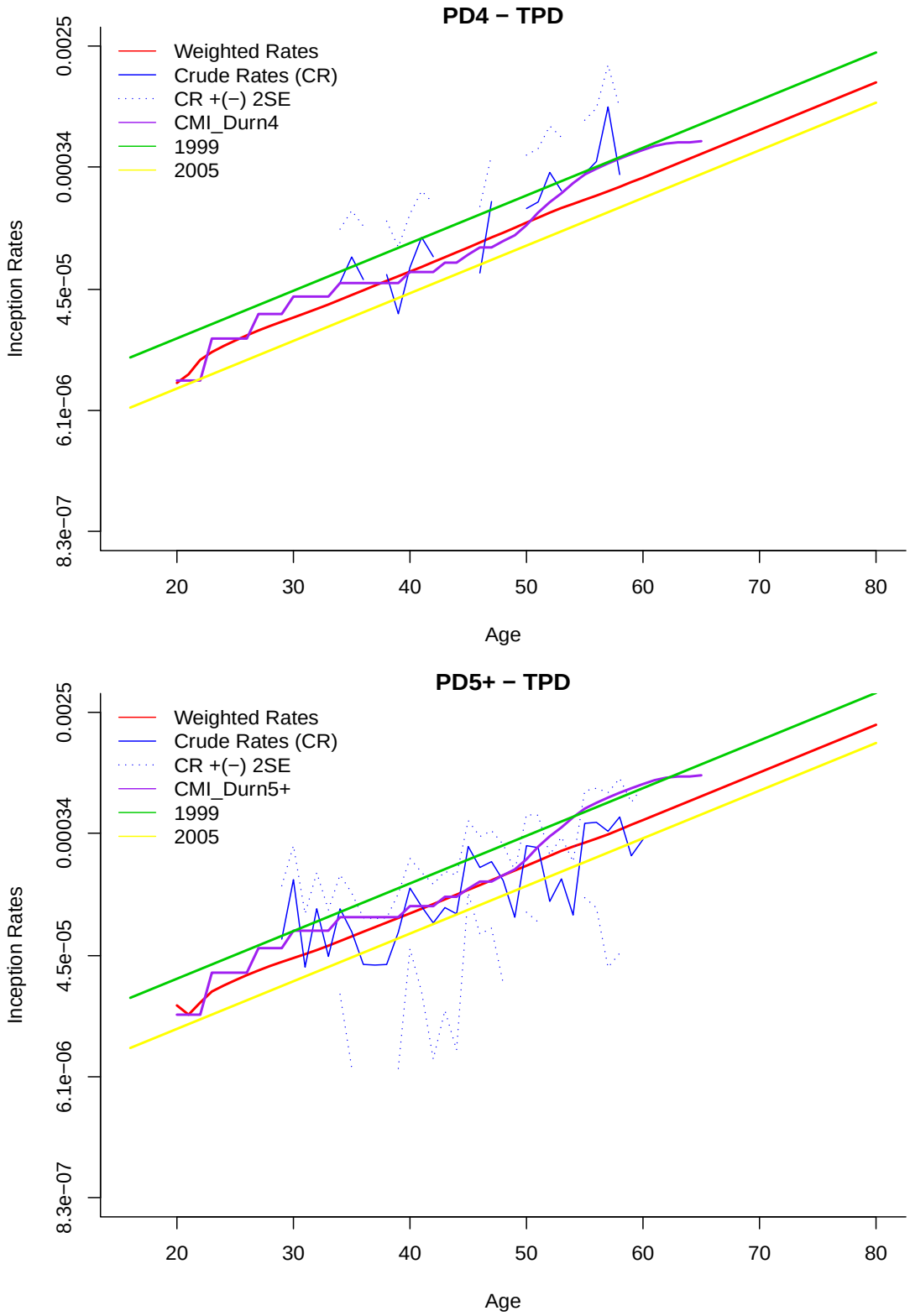


Figure 8.33: Graphs of diagnosis inception rates for TPD for policy durations 4 (PD4) and 5+ (PD5+).

8.3 Comparison of cause-specific rates with the all-causes rates

In this section we give a comparison between the smoothed all-cause rates (calculated in Section 7.3) and the cause-specific rates derived in this chapter. This is useful for verifying our modelling since the cause-specific rates should sum to the all-cause rates, theoretically. This is because the sum of Poisson distributed random variables also follows a Poisson distribution with parameter given by the sum of the individual parameters if the random variables are independent. Here we have two issues. First, we can not guarantee that the numbers of claims for individual causes are independent from each other. For example heart attack and death might not be independent if there are people who suffer from heart attack and do not claim but die afterwards, resulting in a death claim. Secondly, even if we assume independence, the sum of cause-specific intensity rates will not be precisely equal to the all-cause rates, since we adjust exposures for each cause separately. However, we still expect the sum of all-cause rates to be close to the all-cause rates.

One problem we have encountered with these comparisons is non-matching covariates in the best models. To overcome this problem, we took all the covariates appearing in the smoothing functions so far into account, i.e. sex (θ_1), benefit type (θ_2), smoker status (θ_3), year (θ_5), policy duration (θ_7) and office (θ_8). We consider different combinations of these covariates. As an example, the best model obtained in Section 7.3 includes smoker, policy duration and office covariates. So, for different sex or benefit types or years the rates will not change. For TPD, the best model includes year and policy duration, which means these rates will change only with these covariates. On the other hand, for MOT the rates will not change for different risk profiles as the smoothing function only includes an intercept term. Similar comments can be made for all individual causes. Hence, the sum of cause-specific rates is simply the arithmetic sum of individual rates at each age.

Here we consider four different risk profiles, male - non-smokers, female - non-smokers, male - smokers and female - smokers, where we take full accelerated policies, year 2003, policy duration 3 and Office1 as given.

The first risk profile we look at is males - non-smokers. In Figure 8.34, we show contributions of individual causes over age. Here we demonstrate the rates for ages 20 to 65 since most of the policies cease at age 65. The upper grey area shows the difference between the all-cause rates (we use the best model obtained) and the sum of the cause-specific rates. According to the figure, below age 30, death, afterwards cancer, has the biggest contribution to the all-cause rates for male - non–smokers. At older ages death rates increase significantly. A comparison between the all-cause rates and the sum of cause-specific rates is provided in Figure 8.35(a) for the given risk profile. To be able to make a direct comparison with Figure 8.34, the age range 20 to 65 is used in this graph. To give a broader perspective, Figure 8.35(b) shows the logarithm of the difference for ages 16 to 80. Note that in this plot the rates are given in log scale, the y-axis is in original scale. It seems that for this specific risk profile, the sum of cause-specific rates is higher than all-cause rates after age 50. As it is mentioned, we do not expect them to be exactly equal to each other, partly because of the problems with independence or different adjustment factors. However the most important reason is probably the use of different smoothing functions for each of the individual causes and the all-cause rates. The main purpose of this section is to see if there is any particular bias in our modelling. Seeing that the two lines intersect each other, we conclude that there is no particular bias in our modelling.

We have inspected the same risk profile given above for females. Figures 8.36 and 8.37 demonstrate the comparison of all-cause rates and cause-specific rates for female - non-smokers this time. As we are using the best model for the all-cause rates, we have the same rates for males and females since this variable is not used in the model. However when we look at individual causes we see that cancer's contribution to female rates is greater. On the other hand, heart attack and death rates are higher for males. Both Figures 8.35 and 8.37 indicate that we have reasonable models for cause-specific rates. The cause-specific rates cross the all-cause rates at two points, indicating no particular bias in the modelling.

The two other risk profiles we have checked are smokers for two genders. Figures 8.38 and 8.40 show the contribution of individual causes for male - smokers and female -smokers, respectively. For both risk groups, cancer and death contribute most to the inception rates. For males, heart attack rates also add significantly to the rates. The

difference between the smoothed all-cause rates and cause-specific rates is shown in Figure 8.39 for male - smokers. Except for ages 30 to 40, the sum of cause-specific rates is greater than the all-cause rates. On the contrary, for female - smokers the sum of the cause-specific rates is smaller between ages 30 and 70 (see Figure 8.41(b)). Although there are no substantial differences between the two rates, they are larger compared to the non-smoker rates. The reason for this might be having less data for smokers. In both cases (for males and females), the difference is higher at younger and older ages where we have significantly less data.

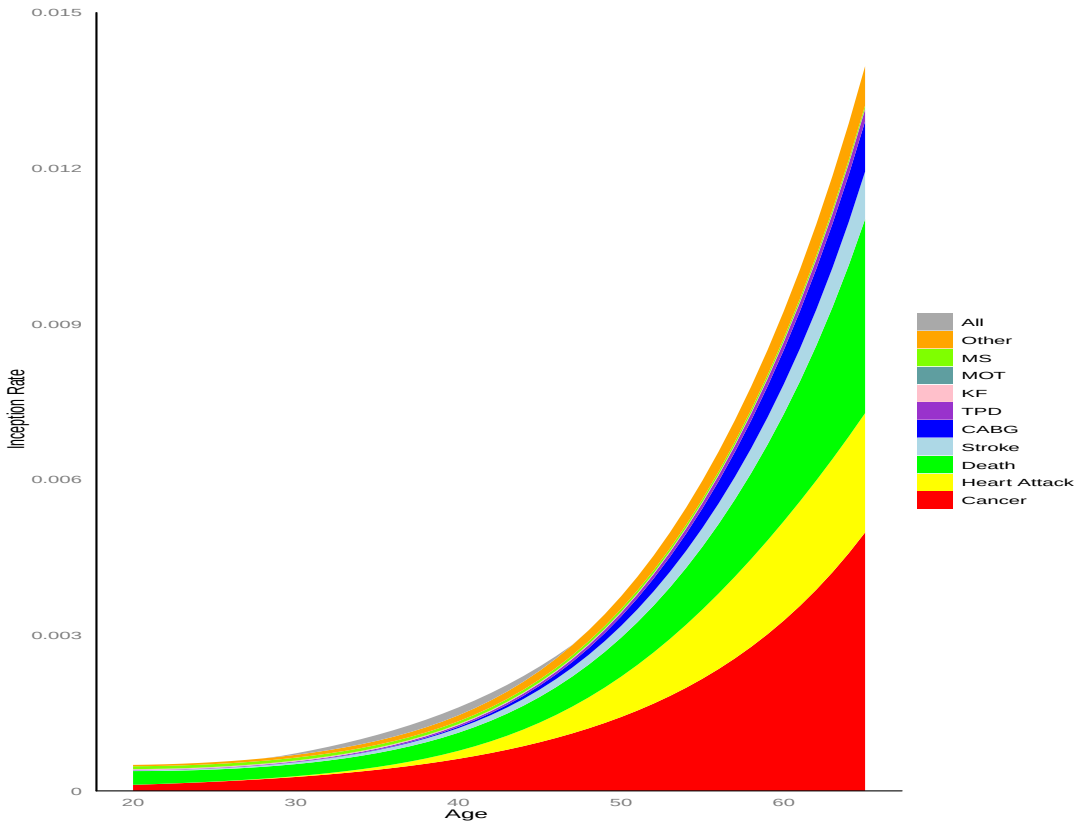
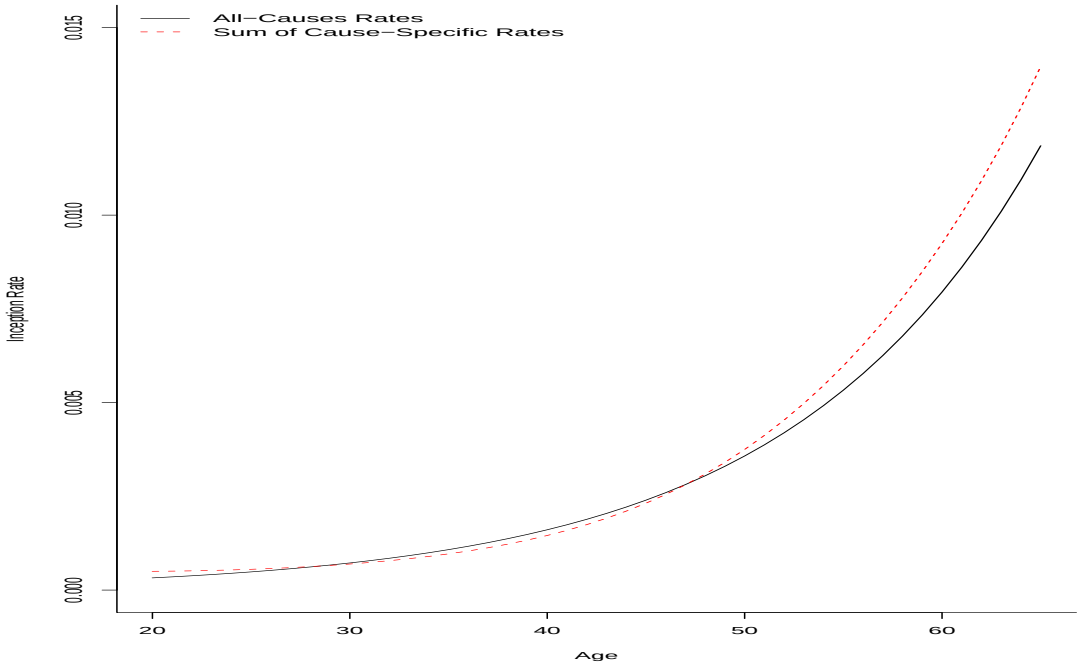
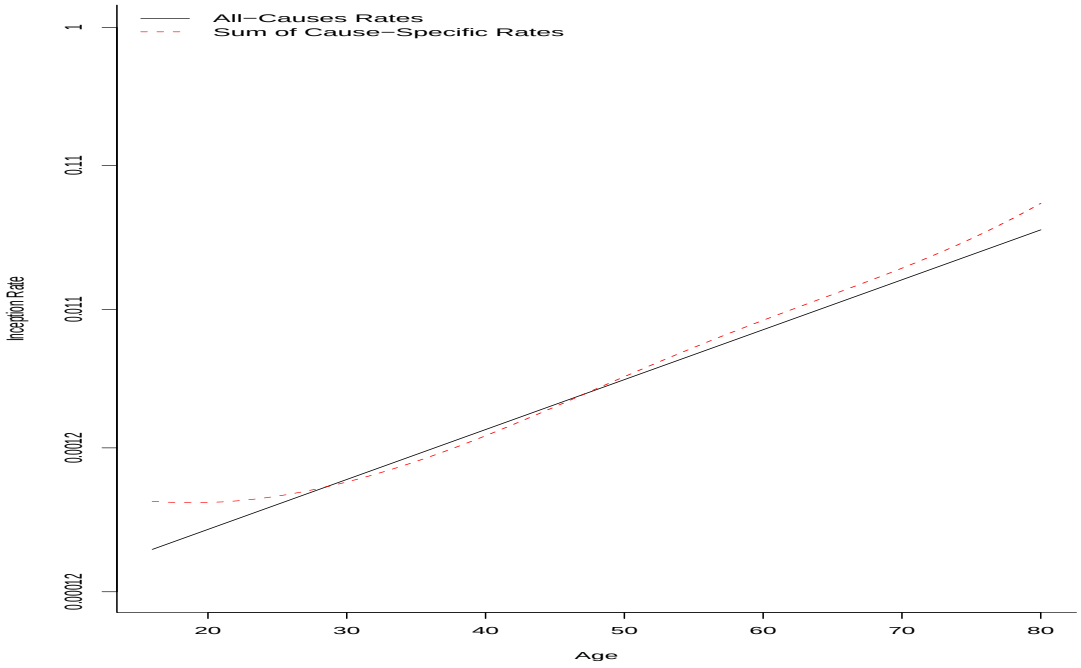


Figure 8.34: Contribution of individual causes for males, full accelerated policies, non-smokers, year 2003, policy durations 3 and Office1.



(a) In actual scale from age 20 to 65.



(b) In log scale from age 16 to 80.

Figure 8.35: Comparison of all-cause rates and summation of cause-specific rates for males, full accelerated policies, non-smokers, year 2003, policy duration 3 and Office1.

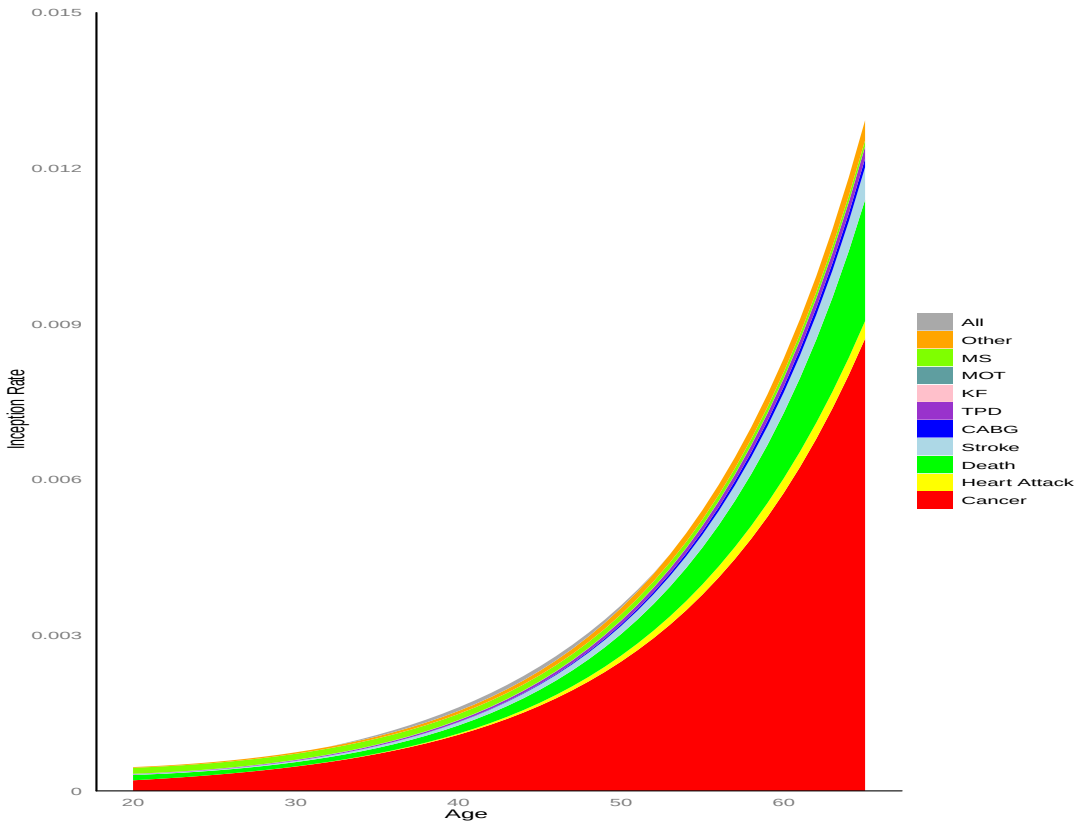
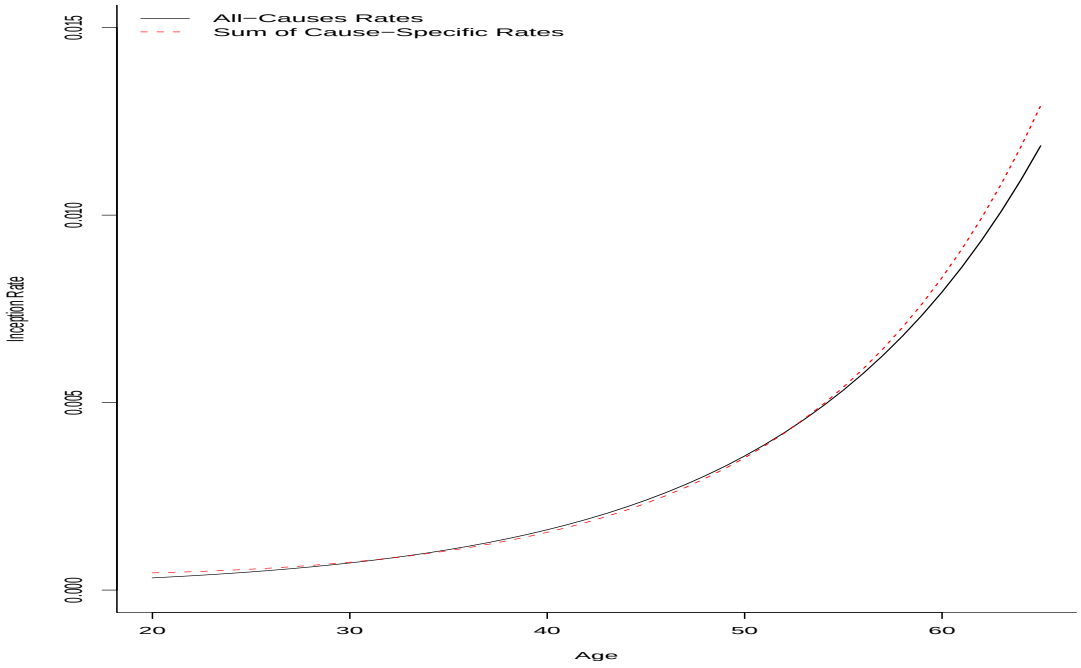
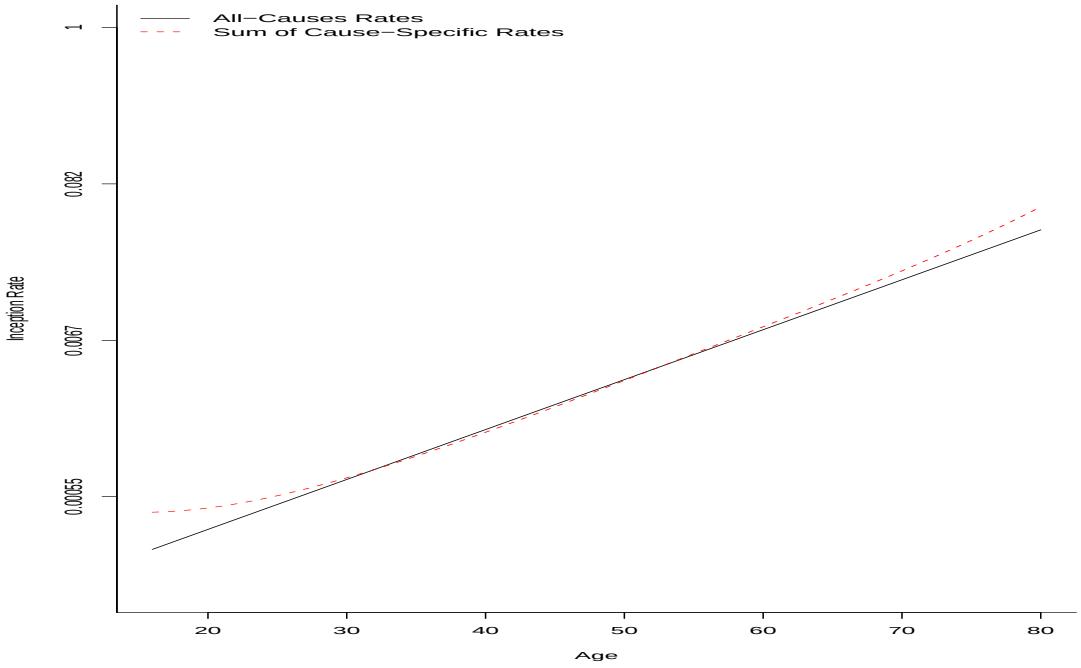


Figure 8.36: Contribution of individual causes for females, full accelerated policies, non-smokers, year 2003, policy durations 3 and Office1.



(a) In actual scale from age 20 to 65.



(b) In log scale from age 16 to 80.

Figure 8.37: Comparison of all-cause rates and summation of cause-specific rates for females, full accelerated policies, non-smokers, year 2003, policy duration 3 and Office1.

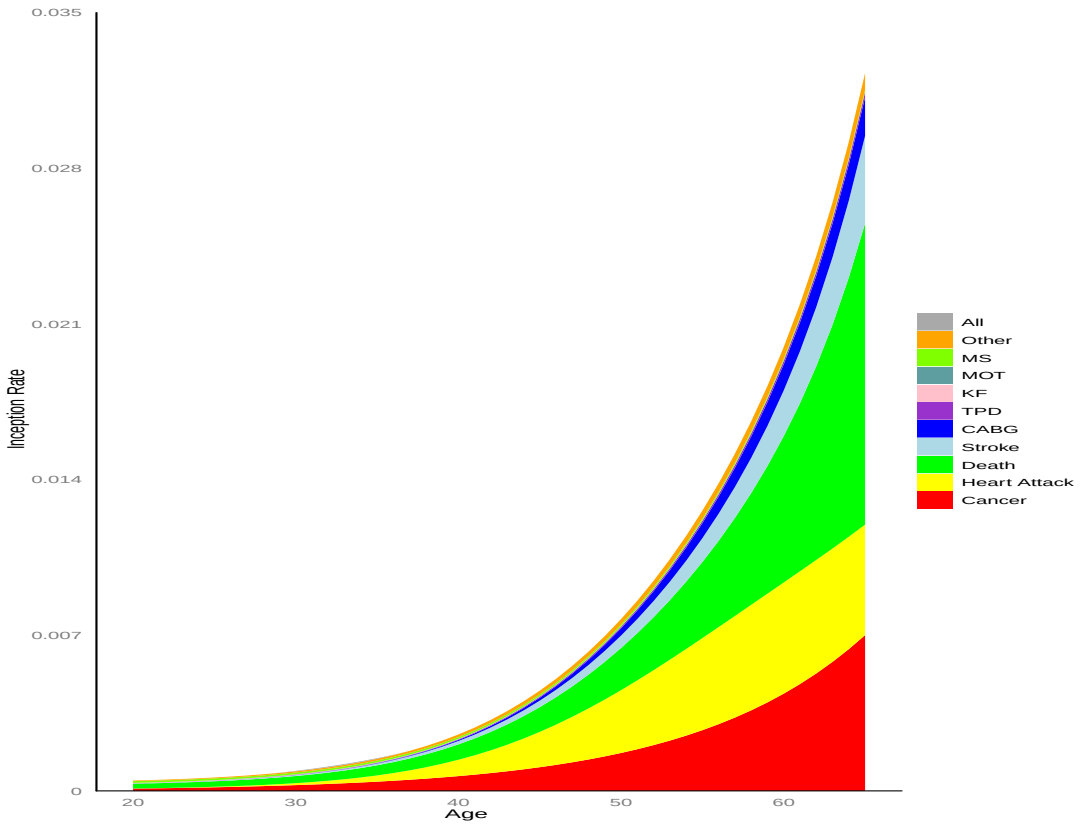
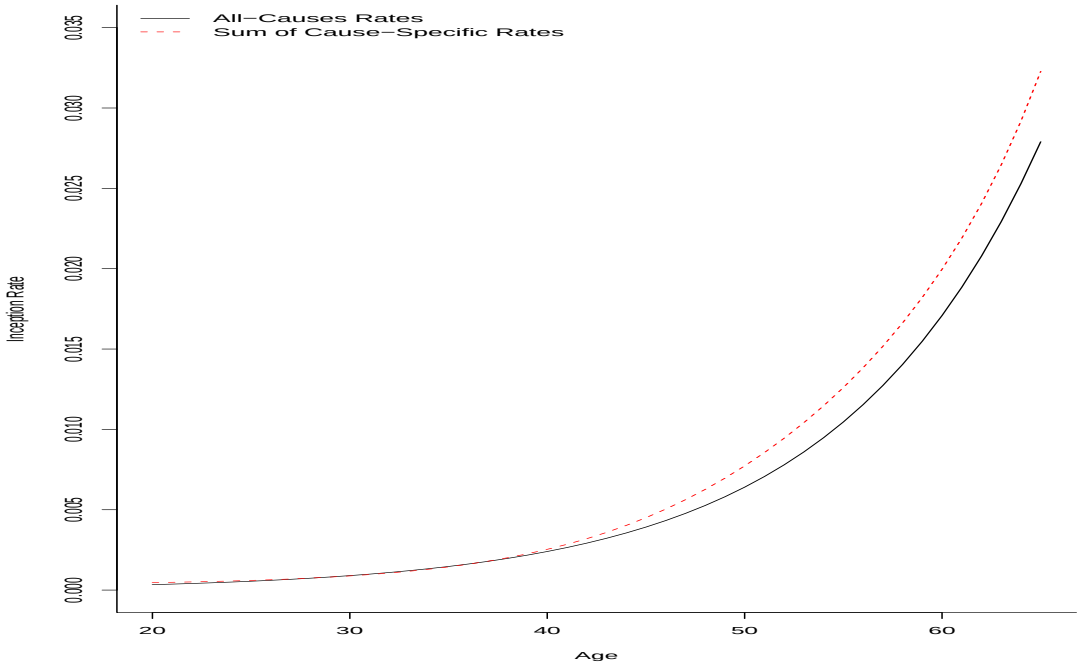
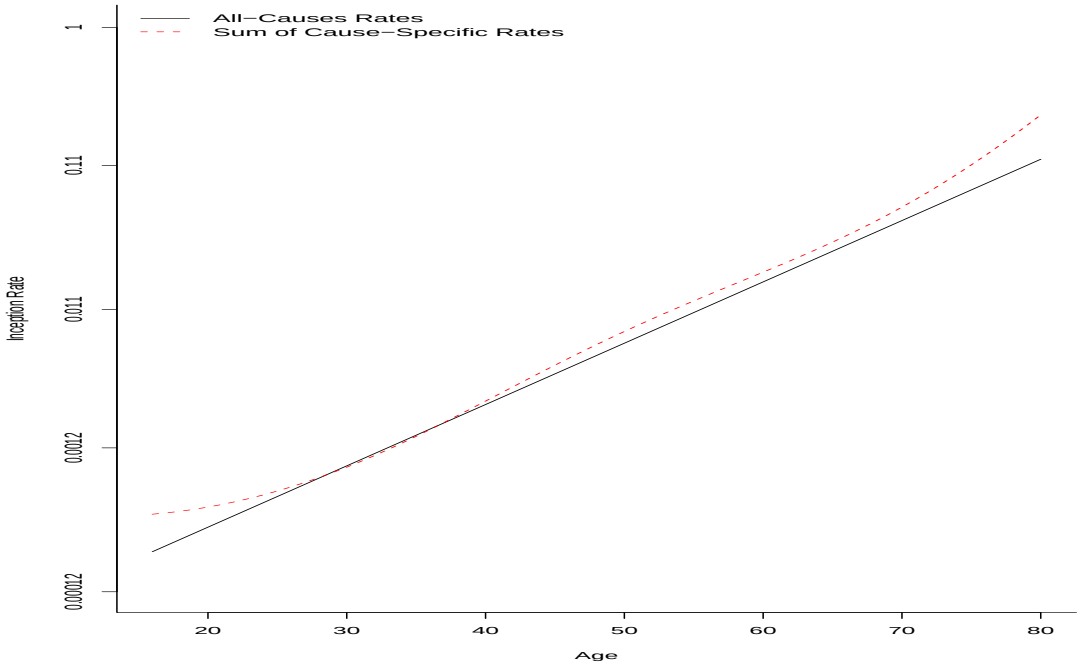


Figure 8.38: Contribution of individual causes for males, full accelerated policies, smokers, year 2003, policy durations 3 and Office1.



(a) In actual scale from age 20 to 65.



(b) In log scale from age 16 to 80.

Figure 8.39: Comparison of all-cause rates and summation of cause-specific rates for males, full accelerated policies, smokers, year 2003, policy durations 3 and Office1.

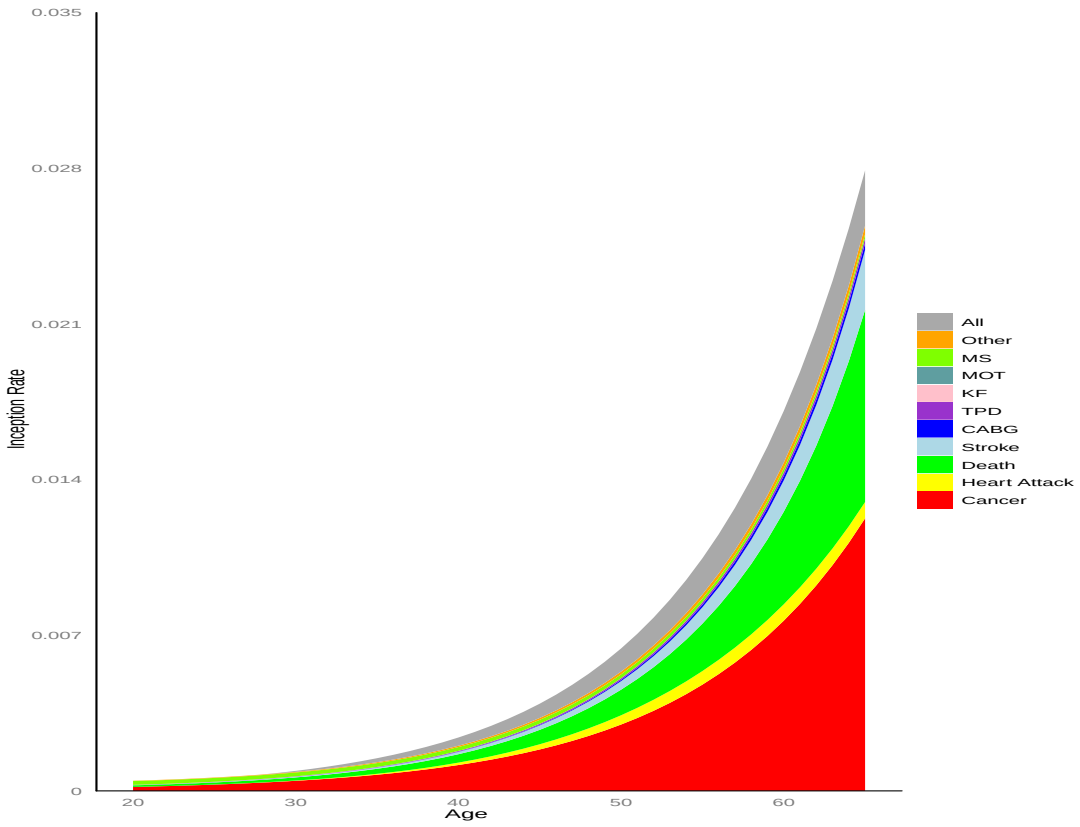
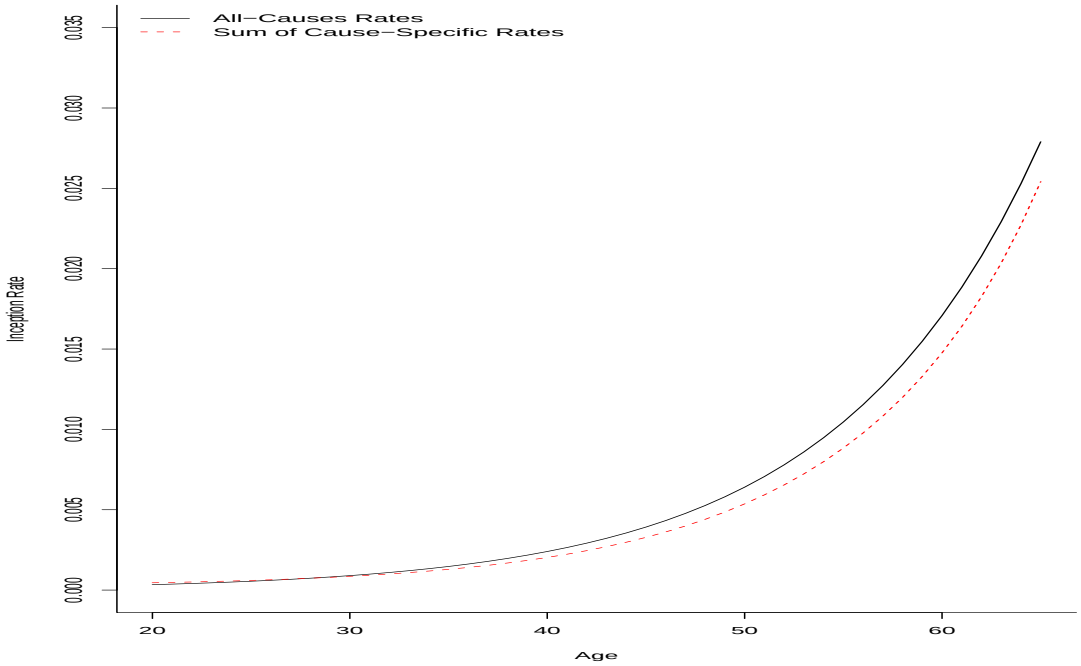
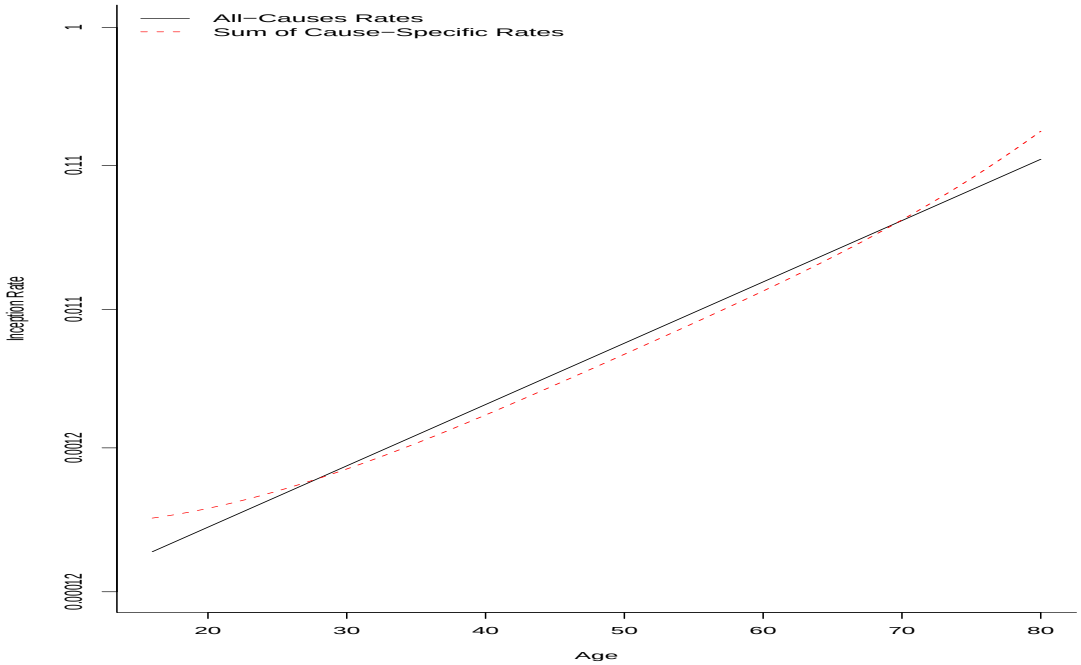


Figure 8.40: Contribution of individual causes for females, full accelerated policies, smokers, year 2003, policy durations 3 and Office1.



(a) In actual scale from age 20 to 65.



(b) In log scale from age 16 to 80.

Figure 8.41: Comparison of all-cause rates and summation of cause-specific rates for females, full accelerated policies, smokers, year 2003, policy duration 3 and Office1.

Chapter 9

Conclusions and Further Research

9.1 Conclusions

One of the main contributions of this thesis is that this is the first statistical model for estimating diagnosis inception rates for CII in the UK, or elsewhere.

The two recent studies by the CMI, WP 43 (2010) and WP 50 (2011), adjust the initial rates for CII by using a base table. Moreover, to analyse specific characteristics, they split the data into subgroups. This approach is limited to providing the inception rates only for the combinations of characteristics where there is a significant amount of data. For all-cause rates, separate rates can be produced for different combinations of sex, smoker status and policy durations. For cause-specific rates, on the other hand, these combinations are generally restricted to males - non–smokers. Apart from these limitations, a main disadvantage of this model is that uncertainty measures (e.g. confidence intervals) can not be provided for inception rate estimates as they are not based on proper statistical modelling. Our approach can provide such measures, as demonstrated for example through the bootstrap confidence intervals in Section 7.5.

In this thesis, we started with modelling the delay between dates of diagnosis and settlement as long delays may distort the results by exposure year. In modelling this delay, both Bayesian and classical methods were used. A three-parameter Burr distribution showed a good fit. However, not being a standard distribution for statistical packages makes it more difficult to use in practice. Therefore we compared the results

with a widely used, standard lognormal model. We showed that this two-parameter distribution gives a poor fit especially in the tail of the observed delay and therefore it is more appropriate to employ a distribution which is more flexible in modelling strongly right-skewed data, such as the Burr distribution.

The effects of settlement year, policy duration, office, death and stroke on the delay were found to be strong when we did not consider missing values or business growth in Chapter 3. When growth rate was introduced in the model in Chapter 4, the strong positive effect of the settlement year disappeared under the Burr model. This is what we expect, as the effect of the settlement year is partly explained by the growth factor as explained in this chapter. Also, with the growth rate in the model, the coefficients for office changed. Again, this is not surprising because we allowed for business growth within each office. However, under the lognormal model, the effect of the settlement year on the delay was not affected. The reason is that the growth rate is mostly affecting the earlier claims since these are given more weight. These earlier claims have longer delay periods because of the nature of the data (data include claims settled between 1999 and 2005). Regarding that, growth rates mostly affect the tail of the distribution which, in this case, was not modelled adequately by the lognormal distribution.

We performed Bayesian variable selection in Chapter 5 in order to obtain the most suitable model to estimate the delay between dates of diagnosis and settlement. As far as we are aware, this is the first study which shows how to perform variable selection under the Bayesian context using a Burr distribution. This is important since the Burr distribution is very appropriate for actuarial analysis as insurance data is generally strongly right-skewed (e.g. claim amounts, number of claims). In model selection, different prior distributions for the model parameters were used. We showed how the Lindley-Bartlett paradox is activated if the priors are not chosen cautiously in Section 5.2. Then, with the empirical and Zellner's g-priors, the best model was determined. While the model excluding age, sex and smoker status was supported without growth rates (in Section 5.2), the model excluding age, sex, smoker status and settlement year was supported when the growth rates were included under the Burr model (in Section 5.3). On the other hand, the model which excludes only sex was found to be the best model with and without growth rates under the lognormal

model. These results were confirmed with Laplace approximation for the Burr model and exact marginal likelihoods for the LN model. We also obtained the same answers from the maximum likelihood based methods as a result of using lots of data and uninformative priors. We note that sex was never found significant under the two models including or excluding the growth rate. However exclusion of age and smoker status was not as certain as sex due to relatively high inclusion probabilities for these variables. In life-related insurance practice, age, sex and smoking status are important policyholder characteristics and when we calculate the inception rates in Chapters 7 and 8 these variables were also included in the analyses. Nevertheless, in Chapter 5 our purpose is to estimate and predict delay in claim settlement, also in the presence of non-recorded dates of diagnosis or settlement. Note that models including age or smoker status were not statistically different from the highest probability model which excludes these variables. However, because of model parsimony we chose to work with the model which excludes these variables. The selected Burr model was used to estimate the missing dates in the data set as it provided a considerably better fit.

Including missing observations in the analysis of the CDD in Chapter 6 changed the posterior densities of some of the model parameters significantly (e.g. some offices, death, TPD). For these variables we have concluded that the missing data mechanism might be systematic rather than random and therefore excluding missing observations from the analyses may lead to bias in the estimation as they may depend on some factors. For example, without including the missing observations, the effect of TPD on the delay is always found to be insignificant with a negative coefficient. However, after taking the missing observations into account this coefficient turned out to be significant with a positive coefficient. This result is reasonable since it is well known that TPD claims take time to be settled. In the CMI's WP 33 (2008, Appendix A), for TPD claims it is mentioned that

> "...(TPD claims) are notoriously difficult to settle, due to delays in establishing the permanence of the disability".

We think that the reason for the ambiguity of this coefficient in our analyses is due to the unclear definition of TPD. There are various different causes of TPD claims including back disorder (24%), mental illness (18%), arthritis (12%) and accidents

(10%) (ABI, 2009). For a TPD claim to be successful, either the claimant should be unable to return to their own or any other occupation, or should not be able to carry out the number of daily living activities as defined in the policy. However, every insurance company applies its own interpretation to assess whether the diagnosis meets the criteria for a TPD claim. This causes problems in CII; approximately 55% of TPD claims are not accepted because the required conditions are not satisfied (see ABI (2009)). Due to these different definitions and the huge range of causes from accidents to mental illnesses, the entered date of diagnosis changes across offices and across TPD claims (e.g. some offices apply a waiting period to some TPD claims and then accept it as the diagnosis date, while some of them accept it as date of diagnosis immediately but wait for some period before they accept it as a claim).

Finally we provided the diagnosis inception rates for all-cause rates (combined) and for 10 individual causes in Chapters 7 and 8, respectively. The observed number of claims was assumed to have a Poisson distribution and the rates were modelled using a similar model to that described in Forfar *et al.* (1988). All-cause rates were modelled under the best model obtained after variable selection in Section 7.3. We concluded that the all-cause diagnosis inception rates mostly depend on age, smoker status, policy duration and office. For general practice we need to provide 'all-office' rates. Here, our approach was weighting the intensity rates with the exposure of each office. At first sight, sex being non-significant in the best model may seem unusual; however we think that the reason is that the high rates for different causes for different sexes cancel each other. In fact, in Chapter 8 it can be seen that for most of the individual causes sex is an important covariate, as expected.

The effects of estimation of missing delays on the inception rates is discussed in Section 7.5. Here it is seen that the inception rates based on other percentiles (from 2.5% to 95%) or the mean of the CDD lie within the confidence interval of the inception rates based on the median of the CDD. This lack of sensitivity is important in practice, since it provides confidence in using estimated inception rates, under possible departures from assumptions used regarding missing values. At this point, we note that this sensitivity analysis could be achieved as a result of obtaining estimates of the CDD under a Bayesian modelling framework. However, obtaining the Bayesian estimates for the inception rates would be computationally very challenging (especially using

WinBUGS) considering the size of the data set.

The sum of individual rates should give the all-cause rates, in theory. Since we used different smoothing models, this will not be satisfied precisely. However to determine if there is any bias in the modelling we compared the sum of the cause-specific rates with the all-cause rates in Chapter 8. As an example, for a specific characteristic, the difference between the sum of cause-specific rates and all-cause rates was compared for four different combinations of sex and smoker status. The results were encouraging as the cause-specific rates and all-cause rates cross each other at least at two points. For the age range where we have most of the data these rates are very close to each other whereas the difference is bigger for younger and older ages due to lack of data.

9.2 Further research

The reason for providing diagnosis inception rates is to estimate the future cash flow of a CI policy and determine the liability of the insurance company more accurately. So, a natural extension of this thesis would be calculation of these expected cash flows, as these can be used in reserving and pricing.

It is important to model the inception rates for older ages since the rates are higher for these ages. One of the approaches in the future might be to compare the incidence rates from our analysis (from insured data) with population statistics. By determining the relationship between these two, we could provide more reliable rates for the ages where we have less data.

Another interesting research topic would be the projection of morbidity rates to future calendar years for different causes for CII.

As mentioned earlier, the claim inception analyses in Chapters 7 and 8 could be performed under a full Bayesian approach so that the entire posterior distribution can be estimated after incorporating observations with missing dates. This would provide proper measures of parameter uncertainty, including corresponding standard errors and credible intervals. We note that this would be computationally very challenging due to the amount of data we have. This problem would require the use of considerably more powerful programming tools than those provided in WinBUGS (e.g. C++).

In this thesis we could not separate different types of cancers. Although we have enough data to analyse some of the cancers on their own (e.g. breast cancer), consultation with the CMI revealed that for many cancer claims cause is recorded as 'cancer - site not specified' by offices and providing any specific cancer rates would underestimate the true rates for specific cancer types. If the site of the cancer can be specified for each claim in the future, then inception rates for specific cancers can be provided.

References

Antonio, K., Beirlant, J. (2008). Issues in Claims Reserving and Credibility: A Semiparametric Approach With Mixed Models. *The Journal of Risk and Insurance*, **75(3)**, 643–676.

Association of British Insurers (2005). *A Guide to Critical Illness Cover*. ABI.

Association of British Insurers (2006). *Statement of Best Practice for Critical Illness Cover*. ABI.

Association of British Insurers (2009). *ABI Statement of Best Practice for Critical Illness Cover. 2009 Review Consultation Paper*. ABI.

Association of British Insurers (2010). *Research into customers understanding of draft Total Permanent Disability headings and definitions*. ABI.

Bartlett, M. (1957). Comment on D.V. Lindley's statistical paradox. *Biometrika*, **44**, 533–534.

Beirlant, J., Goegebeur, Y., Verlaak, R., Vynckier, P. (1998). Burr regression and portfolio segmentation. *Insurance: Mathematics and Economics*, **23**, 231–250.

Beirlant, J., Guillou, A. (2001). Pareto Index Estimation Under Moderate Right Censoring. *Scandinavian Actuarial Journal*, **2**, 111–125.

Beirlant, J., Goegebeur, Y. (2003). Regression with response distributions of Pareto-type. *Computational Statistics & Data Analysis*, **42**, 595–619.

Bowling, A., Bond, M., McKee, D., McClay, M., Banning, A. P., Dudley, N., Elder, A., Martin, A., Blackman, I. (2001). Equity in access to exercise

tolerance testing, coronary angiography, and coronary artery bypass grafting by age, sex and clinical indications. *Heart*, **85(6)**, 680–686.

Bradshaw, P. J., Jamrozik, K., Le, M., Thompson P. L. (2002). Mortality and recurrent cardiac events after CABG: long-term outcomes in a population study. *Heart*, **88**, 488-494.

Cancer Research UK (2010). *Cancer in the UK : July 2010.* Cancer Research UK.

Chatterjee, T., Macdonald, A.S., Waters, H.R. (2009). A model for ischaemic heart disease and stroke I: The model.. *The Annals of Actuarial Science*, **3(1 & 2)**, 45–82.

Continuous Mortality Investigation Committee (2005). *Working Paper 14 - Methodology underlying the 1999-2002 CMI Critical Illness experience investigation.* Institute of Actuaries and Faculty of Actuaries.

Continuous Mortality Investigation Committee (2007). *Working Paper 28 - Progress towards an improved methodology for analysing CMI critical illness experience.* Institute of Actuaries and Faculty of Actuaries.

Continuous Mortality Investigation Committee (2008). *Working Paper 33 - A new methodology for analysing CMI critical illness experience.* Institute of Actuaries and Faculty of Actuaries.

Continuous Mortality Investigation Committee (2010). *Working Paper 43 - CMI critical illness diagnosis rates for accelerated business, 1999-2004.* Institute of Actuaries and Faculty of Actuaries.

Continuous Mortality Investigation Committee (2011). *Working Paper 50 - CMI critical illness diagnosis rates for accelerated business, 2003-2006.* Institute and Faculty of Actuaries.

Critical Illness Trends Research Group (2006). *Exploring the Critical Path.* presented to the Staple Inn Actuarial Society.

Dash, A. C. and Grimshaw, D. L. (1993). Dread Disease Cover - An Actuarial Perspective. *JSS*, **33**, 149–193.

Dellaportas, P., Forster, J., Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, **12**, 27–36.

Dinani, A., Grimshaw, D., Robjohns, N., Somerville, S., Spry, A., Staffurth, J. (2000). *A Critical Review: Report of the Critical Illness Healthcare Study Group.* presented to the Staple Inn Actuarial Society.

Dutang, C., Goulet, V., Pigeon, M. (2008). actuar: An R Package for Actuarial Science. *Journal of Statistical Software*, **25(7)**, 1–37.

Efron, B., Tibshirani R. J. (1993). *An introduction to the bootstrap.* Chapman and Hall.

Forfar, D. O., McCutcheon, J. J., Wilkie, A. D. (1988). On graduation by mathematical formula. *Journal of the Institute of Actuaries*, **115(I)**, 1–149.

Frees, E. W., Valdez, E. A. (2008). Hierarchical Insurance Claims Modeling. *Journal of the American Statistical Association*, **103(484)**, 1457–1469.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (2000). *Bayesian Data Analysis.* Chapman and Hall, USA.

GenRe (2007). *A critical table: Pricing Critical Illness in the UK on a new insured lives table.* presented to the International Actuarial Association Healthcare Section Colloquium in Cape Town.

Gilks, W. R., Richardson, S., Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice.* Chapman and Hall, UK.

Greene, W. H. (1990). *Econometric analysis.* Macmillan.

Hogg, R. V., Klugman, S. A. (1984). *Loss Distributions.* John Wiley & Sons, Inc., USA.

Kass, R. E., Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, **90**, 773–795.

Kass, R. E., Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, **90**, 928–934.

Lindley, D. (1957). A statistical paradox. *Biometrika*, **44**, 187–192.

Macdonald, A. S. (1996a). An Actuarial Survey of Statistical Models for Decrement and Transition Data. I: Multiple State, Binomial and Poisson Models.. *British Actuarial Journal*, **2**, 129–155.

Macdonald, A. S. (1996b). An Actuarial Survey of Statistical Models for Decrement and Transition Data. II: Competing Risks, Non-Parametric and Regression Models.. *British Actuarial Journal*, **2**, 429–448.

Macdonald, A. S. (1996c). An Actuarial Survey of Statistical Models for Decrement and Transition Data. III: Counting Process Models.. *British Actuarial Journal*, **2**, 703–726.

MacDonald, P., Johnstone, D., Rockwood, K. (2000). Coronary artery bypass surgery for elderly patients: is our practice based on evidence or faith?. *CMAJ*, **162**, 1005-1006.

McCullagh, P., Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

Miller, I., Miller, M. (2004). *John E. Freund's Mathematical Statistics with Applications*. Prentice Hall.

NHS (2010). Coronary artery bypass graft [online]. Available from: http://www.nhs.uk/Conditions/Coronary-artery-bypass/Pages/Introduction.aspx. (Accessed 3 February 2011).

Ntzoufras, I. (2002). Gibbs Variable Selection Using BUGS. *Journal of Statistical Software*, **7**, Issue 7.

Ntzoufras, I., Dellaportas, P., Forster, J. J. (2003). Bayesian variable and link determination for generalised linear models. *Journal of Statistical Planning and Inference*, **111**, 165–180.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. John Wiley & Sons, Inc., USA.

R Development Core Team (2009). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Richards, S. J. (2008). Applying survival models to pensioner mortality data. *British Actuarial Journal*, **14(II)**, 257–326.

Schwarz, Gideon E. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6(2)**, 461-464.

Scollnik, D. P. M. (2001). Actuarial Modeling with MCMC and BUGS. *North American Actuarial Journal*, **5(2)**, 96–124.

Society of Actuaries in Ireland Working Party (1994). *Reserving for Critical Illness guarantees*. presented to the Society of Actuaries in Ireland.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B*, **64**, 583–640.

Spiegelhalter, D., Thomas, A., Best, N., Lunn, D. (2003). *WinBUGS User Manual Version 1.4*. MRC Biostatistics Unit, UK.

Stone, P. H., Thompson, B., Anderson, H. V., Kronenberg, M. W., Gibson, R. S., Rogers, W. J., Diver, D. J., Théroux, P., Warnica, J. W., Nasmith, J. B., Kells, C., Kleiman, N., McCabe, C. H., Schactman, M., Knatterud, G. L., Braunwald, E. (1996). Influence of race, sex, and age on management of unstable angina and non-Q-wave myocardial infarction: The TIMI III registry. *JAMA*, **275(14)**, 1104–1012.

Waters, H. R. (1984). An approach to the study of multiple state models. *Journal of the Institute of Actuaries*, **111**, 363–374.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, North-Holland, Amsterdam*, , 233–243.